Universidad Complutense de Madrid

Facultad de Informática

Máster en Internet de las Cosas

Predicción de nubes a corto plazo para una plataforma solar a partir de datos radiométricos

Autor **Álvaro Martín Otero**

Director **Rafael Caballero Roldan**

Colaborador externo **Luis Fernando Zarzalejo Tirado (CIEMAT)**

Trabajo de Fin de Máster

Curso 2017/2018

Convocatoria de septiembre

Calificación: 8,5

Autorización de difusión

Autorización para la difusión del Trabajo Fin de Máster y su depósito en el Repositorio Institucional E-Prints Complutense

El abajo firmante, matriculado en el Máster en Internet de las Cosas de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: "Predicción de nubes a corto plazo para una plataforma solar a partir de datos radiométricos", realizado durante el curso académico 2017-2018 bajo la dirección de Rafael Caballero Roldan [y con la colaboración externa de dirección de Luis Fernando Zarzalejo Tirado (CIEMAT)] en el Departamento de Sistemas Informáticos y Programación, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en internet y garantizar su preservación y acceso a largo plazo.

Firma del alumno

Firma del tutor

Índice general

Índice de figuras

Índice de cuadros

Lista de Acrónimos

IPv6 Internet Procotol version 6

IA Inteligencia artificial

IoT Internet de las cosas

LSTM Long short-term memory

ARIMA Auto Regressive Integrated Moving Average

RMSE Raíz del Error medio cuadrático

CNN Convolutional Neural Networks

RNN Recurrent Neural Networks

Resumen

Predicción de nubes a corto plazo para una plataforma solar a partir de datos radiométricos.

Internet de las cosas es un paradigma que ha revolucionado la conexión entre las personas y los objetos generando en tiempo real una gran cantidad de datos. Debido a esta revolución, diversos campos están viviendo un gran aumento en su utilización, y entre ellos se encuentra el campo de las energías renovables. En concreto, la energía solar está teniendo una velocidad de desarrollo muy acentuada, necesitando nuevas formas de actuar y de gestionar las instalaciones. En este trabajo se aborda el problema de la predicción de radiación global sobre superficie horizontal con alta resolución espacial y temporal (5 minutos) a partir de los datos registrados durante un año en la red radiométrica de alta resolución ubicada en la Plataforma Solar de Almería (PSA-CIEMAT [1]). En particular se muestra un método capaz de predecir el valor de radiación en los siguientes minutos a partir de los valores de los minutos anteriores. El método emplea el tipo de red neuronal recurrente conocido como LSTM, capaz de aprender patrones complejos y predecir el próximo elemento de una serie temporal. Los resultados muestran una mejora apreciable en la precisión del método con respecto a la predicción basada en el último valor conocido.

Palabras clave en Español

- Aprendizaje automático
- Internet de las cosas
- Redes neuronales
- Aprendizaje profundo
- Radiación Solar

[1] http://www.psa.es

Abstract

Short term cloud nowcasting for a solar power station based on historical data.

The Internet of Things is a paradigm that has revolutionized the connection between people and objects, generating a large amount of data in real time. Due to this revolution, diverse fields are experiencing a large increase in their use, and among them is the field of renewable energies. In particular, solar energy has a very high development speed, new ways of acting and managing facilities are needed. This work deals with the problem of the prediction of global radiation on a horizontal surface with high spatial and temporal resolution (5 minutes) from the data recorded during a year in the high resolution radiometric network locate in the Solar Platform of Almería (PSA-CIEMAT [2]). In particular, a method capable of predicting the radiation value in the following minutes from the values of the previous minutes is shown. The method employs the type of recurrent neural red known as LSTM, capable of learning complex patterns and predicting the next element of a time series. The results show an appreciable improvement in the accuracy of the method with respect to the prediction based on the last known value

Keywords in English

- Machine learning
- Internet of things
- Neuronal networks
- Deep learning
- Solar radiation

[2] http://www.psa.es

Capítulo 1

Introducción

1.1. Predicción mediante IoT

Internet de las cosas (Internet of Things, IoT, por sus siglas en inglés) es una revolución, un paradigma, en las conexiones entre las personas y los objetos, proporcionando en tiempo real una gran cantidad de información. IoT consiste en la interconexión de dispositivos a Internet formando redes de diversa topología (en estrella, redes *mesh*, punto a punto...) donde los dispositivos pueden compartir información entre sí para alcanzar un conocimiento superior del entorno que les rodea. El número de dispositivos siempre conectados ha ido creciendo de manera exponencial en los últimos años y se estima que en el año 2020 habrá más de 50 mil millones de dispositivos conectados [1].

Las aplicaciones para las que se diseñan estas redes son muy heterogéneas; en cada campo de aplicación surgen nuevas necesidades apoyadas por el incremento de la integración de procesadores más rápidos, más pequeños, de menor consumo y más asequibles. Además, el desarrollo en el campo de los sensores permite recopilar una gran cantidad de información ya que estos mismos pueden ser ubicados en múltiples emplazamientos. Del mismo modo, el incremento del número de las infraestructuras de comunicaciones inalámbricas y de interfaces de comunicación de bajo consumo, apoyado por la expansión en el espacio de direcciones que ofrece IPv6, está propiciando la expansión de estos dispositivos siempre conectados.

Muchos de los datos recogidos por estos dispositivos interconectados se caracterizan por tener un gran volumen, ya que se produce un envío constante desde muchos sensores, una gran variedad, debido a la heterogeneidad de las redes formadas, y una gran velocidad de generación de estos datos. Estas tres cualidades forman la definición del *big data*.
Big data es un término que hace referencia a grandes cantidades de datos, ya sean estos estructurados, no estructurados o semiestructurados, que las técnicas convencionales de procesamiento, gestión y almacenamiento no son capaces de administrar. Para que un conjunto de datos sea considerado *Big data* debe cumplir al menos las 3 cualidades conocidas como las 3 Vs: volumen de datos, velocidad de generación y variedad de datos recopilados. Existen también definiciones alternativas que extienden de 3 hasta 5 Vs, a las que añaden la veracidad, los datos recibidos no deben ser incompletos o incorrectos, y el valor, que los datos puedan convertirse en información.

Este gran aumento en el campo del *big data* genera la necesidad de nuevas técnicas de análisis de datos debido a que la información almacenada debe ser tratada para convertirla en información. El campo encargado de diseñar los algoritmos para sacar valor a estos datos es el de la inteligencia artificial (IA), pero más en concreto la rama de *machine learning*.

Según [2], "llamamos *machine learning* o aprendizaje automático a aquellos programas que mejoran su eficiencia al realizar ciertas tareas tras pasar por un proceso de aprendizaje". Su objetivo final es tratar de conseguir que los computadores puedan emular el comportamiento humano. Esta técnica de IA tiene un gran número de aplicaciones a diferentes campos, y entre ellos se encuentra el de las energías renovables.

Las energías renovables son fuentes de energía limpias, inagotables y crecientemente competitivas. El desarrollo de estas energías sostenibles es imprescindible para combatir contra el cambio climático y la problemática medioambiental asociada a los residuos y emisiones que producen los sistemas de producción convencionales. Entre estas energías renovables se encuentra la energía solar, la cual está consiguiendo un gran aumento de producción y de instalaciones de plantas solares.

Exactamente en este Trabajo de Fin de Máster se desarrolla una metodología de estudio para la predicción de radiación solar global a partir de los datos previamente registrados durante un año en la red radiométrica de la Plataforma Solar de Almería (PSA-CIEMAT, [1]) utilizando técnicas de análisis de datos.

1.2. Caso de estudio: predicción de radiación solar

Uno de los principales problemas relacionados con las energías es la sostenibilidad [3]. Este término implica poder satisfacer las necesidades actuales pero sin comprometer las que ocurrirán. Para poder garantizar esta preservación de los recursos es necesario fomentar e incrementar el uso de las energías renovables, evitando así la gran cantidad de residuos y contaminación producida por los sistemas convencionales de producción. Aprovechando el impulso que supone los beneficios que las energías renovables aportan, se está realizando un gran esfuerzo investigador para generar innovaciones tecnológicas en este tipo de energía, en concreto en la energía solar, cuyos componentes han conseguido disminuir su vulnerabilidad frente al paso del tiempo disminuyendo así su coste en grandes cantidades.

La forma de conseguir energía en las plantas solares es mediante el aprovechamiento de la radiación solar, por tanto, es importante ser capaz de dar una estimación precisa de la cantidad de energía que se producirá en un instante concreto. Al no disponer de esta información se pueden producir problemas de abastecimiento si no se realiza una correcta gestión.

Actualmente existen plataformas que realizan predicciones meteorológicas [2], pero no todas ellas son capaces de realizar estas predicciones en intervalos de tiempo reducido. Por tanto, existe

[1] http://www.psa.es
[2] http://www.aemet.es/es/portada

una necesidad de obtener estas predicciones de la radiación solar con un horizonte temporal reducido.
Debido a todo esto, en este Trabajo de Fin de Máster, se realizará una predicción de la radiación solar en un horizonte temporal reducido mediante algoritmos de aprendizaje automático a partir de los datos de radiación históricos.

1.3. Motivación y propósito del estudio

La capacidad de predecir la radiación solar afecta directamente a los gestores de las plantas solares, que requieren de esa previsión para participar en el mercado de la energía y planificar las operaciones de mantenimiento. En definitiva, la capacidad de predecir el recurso solar disponible es crítica para los operadores de plantas solares, ya que esto puede afectar la gestión de la planta y, en consecuencia, la generación de electricidad resultante. Todos estos problemas pueden suponer grandes perdidas tanto energéticas como económicas, por tanto hay que tratar de optimizar al máximo la producción.

Este estudio se centrará en la predicción a corto plazo de la radiación solar en el caso concreto de la Plataforma Solar de Almería mediante los datos de radiación obtenidos de una red de radiómetros solares durante un año con una frecuencia de un minuto.

En el Capítulo 3 se comentará la situación actual de las energías renovables en el IoT profundizando más concretamente en la solar, además del uso de *Deep learning* para la predicción. También se explicarán las tecnologías que se utilizarán en este trabajo además de algunos conceptos de geometría solar. A continuación, en el Capítulo 4 se expondrá cuál ha sido la metodología utilizada en el trabajo, explicando que algoritmos se utilizarán para realizar las predicciones. Tras ello en el Capítulo 5 se profundizará en cómo se ha realizado el procesado de los datos para posteriormente utilizarlos en la predicción. En el Capítulo 6 se mostrarán los resultados obtenidos para finalmente en el Capítulo 7 mostrar las conclusiones obtenidas y las futuras líneas de investigación posibles partiendo de este estudio en el Capítulo 8.

1.3.1. Objetivos de este proyecto

El objetivo final de este proyecto es conseguir una predicción a corto plazo de la radiación solar. Para conseguir este objetivo vamos a desglosar cuáles serán las diferentes tareas a realizar:

***Preparación de los datos para posterior uso*:** Los datos que se utilizarán en este proyecto han sido almacenados según realizaba las lecturas cada una de las estaciones meteorológicas, por tanto, es necesario filtrarlos y procesarlos antes de comenzar con la predicción.

***Estudio previo del conjunto de datos radiométrico*:** Para poder trabajar con él, primero hay que comprender el significado de cada una de las medidas ofrecidas por las estaciones además de analizar cuáles son los parámetros que suponen variaciones en las mismas.

***Estudio de los algoritmos de predicción utilizados*:** Analizar las ventajas y desventajas de los diferentes algoritmos posibles para realizar la predicción y decidir cuáles serán los aplicados en este trabajo.

***Predicción utilizando los datos*:** Una vez decidido qué algoritmo utilizar, se utilizará con los datos procesados previamente para comparar qué resultado ofrece cada uno de ellos.

***Agrupación de los datos*:** Analizar las grupos obtenidos tras agrupar las mediciones similares mediante algoritmos de clustering.

***Análisis de las predicciones obtenidas*:** Examinar los diferentes resultados conseguidos con los algoritmos de predicción y exponer las conclusiones obtenidas con cada uno.

1.3.2. Fases del trabajo

Con el fin de cumplir los objetivos arriba mencionados, el proyecto se divide en las siguientes fases:

1. Se realizará un estudio previo de los datos para encontrar patrones y aplicar ciertas correcciones como eliminar datos ausentes o retrasos en las lecturas del dato. En esta fase se representarán los datos gráficamente e incluso se realizará un vídeo que muestra cómo cambia la radiación según cambia el tiempo.

2. Examinar las diferentes posibilidades que existen para realizar predicción con el tipo de datos que poseemos y estudiar las ventajas y desventajas de cada uno de ellos.

3. En paralelo al estudio de algoritmos de predicción se realizarán métodos de agrupación con los cuáles se estudiaran los grupos obtenidos.

4. Aplicar las diferentes técnicas de predicción.

5. Comparar las métricas obtenidas por cada una de las predicciones realizadas.

Capítulo 2

Introduction

2.1. Prediction through IoT

The Internet of Things (IoT) is a revolution, a paradigm, in the connections between people and objects, providing in real time a big amount of information. IoT consists of the interconnection of devices to the Internet leading to many different topologies of networks (star, mesh, point to point...) where the devices can share information to each other to reach a superior knowledge of the environment that surrounds them. The number of connected devices has grown exponentially in recent years and it is estimated that by 2020 there will be more than 50 billion connected devices[1].

The applications for which these networks are designed are very heterogeneous. In each field of application new needs arise supported by the increase of the integration of faster, smaller, and less consuming processors. Also the development in the field of sensors allows you to collect a lot of information since these can be located in multiple locations. Similarly an increase in the number of infrastructures of wireless communications with low communication interfaces consumption, supported by expansion in the address space offered by IPV6, is promoting the expansion of these devices always connected.

Much of the data collected by these interconnected devices are characterized by having a large volume, since, thanks to the heterogeneity of the formed networks and high-speed generation of these data there is a constant shipment from many sensors with a great variety. These three qualities are the definition of big data.
Big data is a term that refers to large amounts of data, whether structured, unstructured or semi-structured, that conventional processing, management and storage techniques are not able to manage. For a data set to be considered Big data must comply at least the 3 qualities known as the 3 Vs: data volume, generation speed and variety of data collected. There are also alternative definitions that extend from 3 to 5 Vs, which they add truthfulness, the data received must not be incomplete or incorrect, and the value, the data can be converted into information. This great increase in the field of big data generates the need for new techniques of data analysis because the stored information must be treated to convert it into information. The field responsible for designing the

algorithms to extract value from this data is artificial intelligence (AI), but more specifically the machine learning branch.
According [2], "the field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience".Its ultimate goal is to try that computer can emulate human behaviour. This technique of AI has a large number of applications to different fields, among them is the renewable energy sources.

Renewable energies are clean, endless, and increasingly competitive sources of energy. The development of these sustainable energies is essential to fight against change climate and environmental issues associated with residues and emissions that the conventional production systems produce. These renewable energies include solar power, which is getting a large increase in production and installations of solar plants.

For this reason, in this Master thesis, it will be developed a study methodology for the prediction of global solar radiation data previously obtained for a year of a radiometric network located in a radiometric network of Platform Solar in Almería (PSA-CIEMAT [1]) using data analysis techniques.

2.2. Study case: Prediction of solar radiation

One of the main problems related to energy is sustainability [3]. This term implies to meet current needs without compromising those that will occur. To be able to guarantee this resources preservation is necessary to promote and increase the use of renewable energy, avoiding a large amount of waste and pollution produced by conventional production systems. Building on the momentum that the benefits renewable energies provide, is being made a great research effort for generating technological innovations in this type of energy, in particular in solar energy, whose components have managed to reduce its vulnerability to the passage of time decreasing in this way its cost in large quantities.

The way to get energy in solar plants is through the use of the solar radiation, so it is important to be able to give an accurate estimate of the amount of energy that would be produced at a particular moment. Not having this information can produce supply problems if there is not a proper management.

Currently, there are platforms that perform weather forecasts [2], but none of these is able to predict in short time intervals, therefore, there is a need to be able to predict with a reduced time horizon which will be the solar radiation that will occur.
Due to all this, in this Master thesis, a prediction of solar radiation on a time horizon reduced by automatic learning algorithms from historical radiation data will be made.

[1] http://www.psa.es
[2] http://www.aemet.es/es/portada

2.3. Motivation and purpose of the study

The ability to predict the solar radiation directly affects to the managers of solar plants, requiring that forecast to participate in the energy market and plan the maintenance operations. In short, the ability to predict the solar resource available is critical for operators of solar plants, since this may affect the management of the plant and, as a result, the resulting electricity generation.
All these problems con suppose deep loss both energetic and economical, hence a maximum optimization of the production is required.

This study will focus on the short-term prediction of solar radiation in the specific case of the Almeria solar platform through the radiation data obtained from a radiometric sensor network for one year with a frequency of one minute.

In Capítulo 3 the current situation of renewable energies, more specifically solar energy, in the IoT, as well as the use of Deep Learning for prediction, will be discussed. It will also explain the technologies that will be used in this work in addition to solar geometry. The following in Capítulo 4 the methodology which has been used in the work will be described, explaining which algorithms will be used to make the predictions. After that, in Capítulo 5, it will be explained in details how the processing of the data has been done and then used in the prediction. Furthermore, Capítulo 6 shows the results obtained for a later enumeration of conclusions that will take place in Capítulo 7, and finally, in Capítulo 8 it will be depicted the future lines of possible research based on the conclusions of this study.

2.3.1. Objectives of this project

The final objective of this project is to achieve a short-term prediction of solar radiation. To achieve this goal, we will break down the different tasks to be carried out:

***Preparing the data for later use*:** The data that will be used in this project have been stored according to the readings of each meteorological stations, so it is necessary to filter and process them before starting the prediction.

***Previous study of the radiometric dataset*:** To be able to work with the dataset, first we must understand the meaning of each measure read by the stations, in addition to analyzing which are the parameters that suppose variations in them.

***Study of prediction algorithms used*:** Analyze the pros and cons of the different possible algorithms to make the prediction and decide which will be applied in this work.

***Data prediction*:** Once you have decided which algorithm to use, it will be used with the previously processed dataset to compare the results offered by each of them.

***Clustering the data*:** Analyze the groups obtained after grouping similar measurements using clustering algorithms.

***Analysis of the predictions obtained*:** Examine the different results obtained with the prediction algorithms and present the conclusions obtained with each one.

2.3.2. Phases of Work

In order to meet the above objectives, the project is divided into the following phases:

1. A preliminary study of the data will be conducted to find patterns and apply certain corrections such as eliminating missing data or delays in the readings of the data. In this phase the data will be represented graphically and there will even be a video showing how the radiation changes as time changes.
2. With the filtered data, a study of the different prediction algorithms will be carried out and the pros and cons of each one.
3. In parallel, the data will be grouped and conclusions were drawn.
4. Apply the different prediction techniques.
5. Compare the metrics obtained by each of the predictions made.

Capítulo 3

Estado del arte

3.1. Energías renovables en IoT

El sector de las energías renovables está viviendo un gran aumento en su utilización respecto a las convencionales, aumentando así la velocidad a la que se está desarrollando. Debido a este crecimiento en las energías renovables, la industria se encuentra con el reto de ser capaz de procesar la información recopilada por las numerosas instalaciones generadoras en mayor cantidad y de forma más distribuida. Este aumento en el desarrollo de las energías renovables está especialmente acentuado en el campo de la energía solar, la cual ha disminuido considerablemente sus costes en los últimos años.

Gracias a la integración con el IoT, estos retos pueden tratarse de manera más sencilla y eficaz utilizando redes de sensores para recopilar la información y posteriormente procesarla, permitiendo así mejorar la relación entre la producción y la demanda, optimizar el rendimiento de los paneles solares y realizar mantenimiento predictivo.

Uno de los pilares del IoT es la recepción de datos, que pueden ser utilizados para la predicción mediante *machine learning*. El gran avance que se está produciendo en el campo de la sensórica permite realizar numerosos tipos de despliegues interconectando diferentes tipos de sensores, facilitando así la recolección de datos para su posterior utilización además de mejorar la eficiencia en la gestión de las redes inteligentes y la energía distribuida.

3.2. Machine learning

Como se explicó en el Capítulo 1, machine learning [4] es una disciplina de la inteligencia artificial que permite a los sistemas aprender de forma automática mediante la inducción del conocimiento. Por aprender entendemos generar comportamientos a partir de información suministrada previamente en forma de ejemplos. El machine learning explora el estudio de algoritmos que pueden aprender y hacer predicciones sobre datos. Mediante la creación de un modelo utilizando los ejemplos proporcionados como entrada son capaces de realizar predicciones o tomar decisiones.

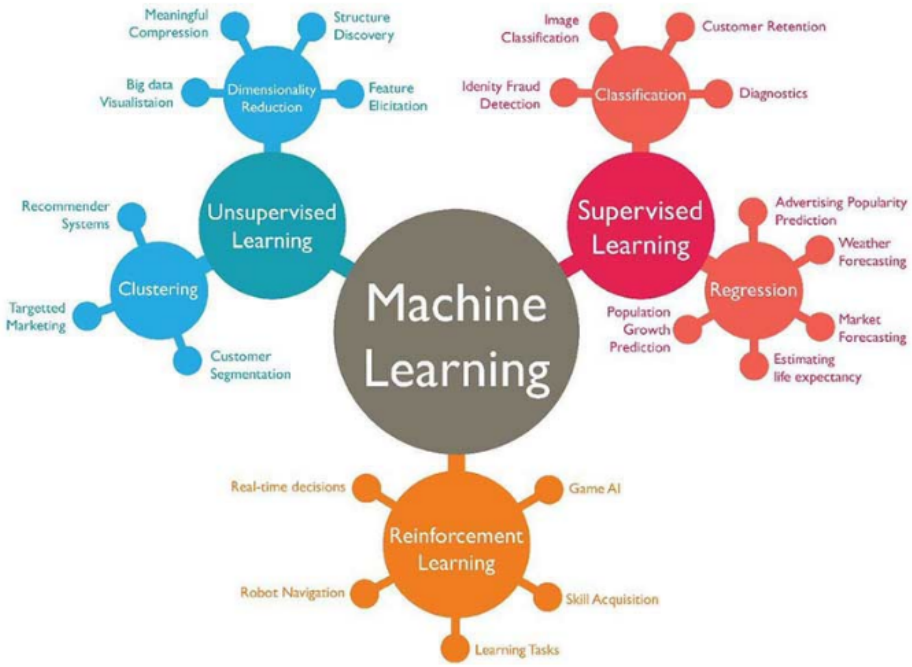


Figura 3.1: La imagen muestra los diferentes tipos en los que se pueden clasificar los algoritmos de *machine learning* y algún ejemplo de uso que tiene cada uno de ellos [2].

Los algoritmos de machine learning pueden clasificarse en diversos grupos dependiendo de si existe o no una señal de entrada:

***Aprendizaje supervisado*:** consiste en deducir una función a partir de unos datos de entrada y unas salidas esperadas. Es decir, establece una correspondencia entre las entradas y las salidas deseadas del sistema. Una vez terminado el proceso de aprendizaje, la función resultante debe ser capaz de predecir el valor correspondiente a la entrada proporcionada. Los datos de entrada comúnmente son pares de objetos, el valor de entrada y la etiqueta, que nos indica el valor que debe producir la salida. Se emplea tanto en labores de clasificación como en herramientas predictivas.

Algoritmos comunes de aprendizaje supervisado:

- Naive Bayes
- Redes neuronales
- Regresión lineal

***Aprendizaje no supervisado*:** es un método ajustado a las observaciones. Al contrario del aprendizaje supervisado, el conjunto de datos de entrada no proporciona el valor que debe generar la salida, es decir, no están clasificados. Por tanto, en el aprendizaje no supervisado se tratan los datos de entrada como variables aleatorias, para construir un modelo de densidad con el

[2]Fuente: https://www.datasciencecentral.com

conjunto de datos proporcionado. Se suele utilizar para el agrupamiento de datos.
Algoritmos comunes de aprendizaje no supervisado:

- k-means
- Reglas de asociación

***Aprendizaje semisupervisado*:** este método es una mezcla de los dos aprendizajes explicados previamente. Habitualmente utiliza una pequeña cantidad de datos etiquetados y una gran cantidad de datos sin etiqueta.

***Aprendizaje por refuerzo*:** este tipo de aprendizaje recibe *feedback* de las decisiones que va tomando, obteniendo la valoración de la idoneidad de su respuesta. Cuando ofrece respuestas correctas el funcionamiento es similar el supervisado, en ambos casos reciben información de la salida correcta. En cambio, cuando la respuesta es incorrecta en el supervisado se le indica la respuesta que debería haber dado, en cambio, en el de refuerzo se le informa que ha sido incorrecto y se cuantifica este error.
Algoritmos comunes de aprendizaje por refuerzo:

- Q-Learning
- Deep Adversarial Networks

Según el tipo de resultado que queramos obtener del aprendizaje, podemos categorizar *machine learning* en:

***Clasificación*:** aprendizaje supervisado en el que los datos de entrada se dividen en dos o más clases. Para cada una de las entradas se debe asignar una etiqueta que lo asocie a una de las clases en las que hemos dividido el conjunto de datos. Se busca conseguir una función que sea capaz de convertir nuestras entradas a las variables de salida discretas. En este tipo de aprendizaje queremos predecir una etiqueta discreta.

***Regresión*:** consiste en encontrar la relación existente entre los conjuntos de entrada. Nos permite observar como cambia el valor de la variable dependiente cuando se modifica la independiente. Se busca conseguir una función que sea capaz de convertir nuestras entradas a una variable de salida continua. En este tipo de aprendizaje queremos predecir un valor continuo.

***Agrupación*:** aprendizaje no supervisado que al igual que la clasificación se divide el conjunto de entrada en grupos, pero en este caso no se conocen de antemano. Aquí el objetivo es ser capaz de ordenar los datos de entradas en grupos o clúster, de modo que el grado de asociación sea fuerte entre los miembros de un mismo grupo y débil entre los miembros de los demás grupos.

3.2.1. Deep learning

Deep Learning es un conjunto de algoritmos de machine learning basados en redes neuronales artificiales para el aprendizaje automático. Pertenece a los métodos de aprendizaje supervisado. Estas redes están formadas por múltiples capas con unidades de procesamiento no lineal para obtener características y realizar transformaciones. La capa de entrada recibe la señal de entrada, las capas ocultas procesan las salidas de las capas anteriores sucesivamente hasta proporcionar una salida. Cada una de estas capas está formada por unidades neuronales, las cuales operan empleando funciones suma, cuya salida está conectada con la entrada de la siguiente. Mediante esta estructura es capaz de realizar diferentes niveles de abstracción.

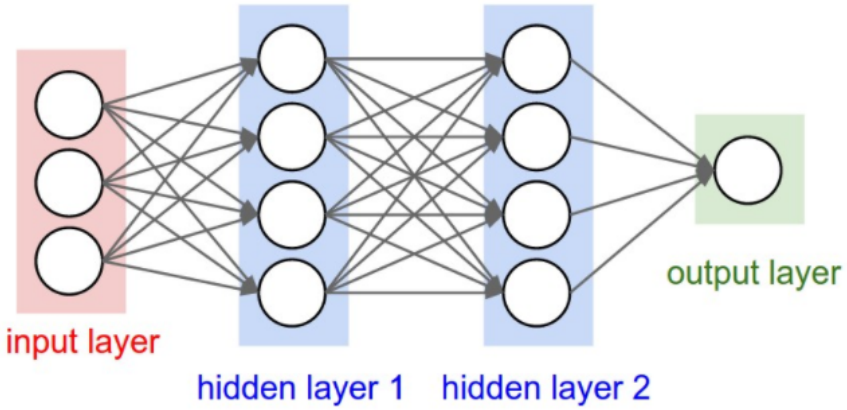


Figura 3.2: Ejemplo de una red neuronal artificial con dos capas intermedias de cuatro nodos cada una. Fuente:[5].

Cada capa está formada por nodos, los cuales son el lugar donde se realiza el cómputo, inspirados en las neuronas del cerebro humano, que se disparan cuando reciben suficientes estímulos. Un nodo combina la entrada con una serie de pesos, consiguiendo amplificar o amortiguar esa señal, para así poder poner más importancia a unas u otras entradas. Tras ello se suman todos estos valores resultantes de la combinación y se pasan por una función de activación, la cual indica si esa señal debe pasar a través de la red o no. Estas funciones son binarias (se activa o no) y las que se utilizan comúnmente son la función signo, la función semilineal y la función sigmoidea. Ajustar bien los pesos es lo que permite indicar la importancia que le damos a la característica que recibimos de entrada respecto a como la red clasifica o genera grupos. En la Figura 3.3 estos pesos están representados por w_i, siendo i el número de la entrada correspondiente a dicho peso.

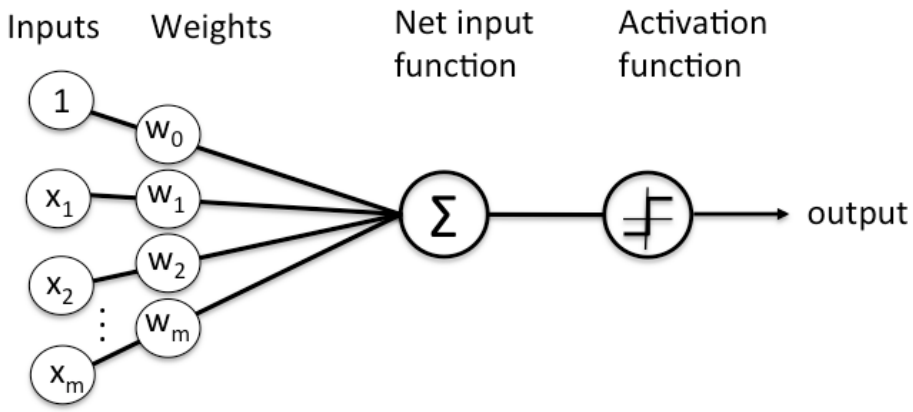


Figura 3.3: Ejemplo de un nodo de una red neuronal, en la que se pueden ver las diferentes fases que se producen. Fuente:[6]

Este concepto de red neuronal se remonta a hace más de medio siglo, pero ahora está en auge debido a la gran cantidad de datos que se genera con el IoT y el avance que se ha vivido en la capacidad de almacenamiento y de cómputo.

Una red neuronal de grandes dimensiones, dispone de muchas capas intermedias y cada una con un gran número de nodos, lo que supone una cantidad exponencial de parámetros para recibir. Esto quiere decir que es necesaria una gran capacidad de cómputo, y que sin suficientes datos no se podrá realizar el proceso de aprendizaje de forma óptima.

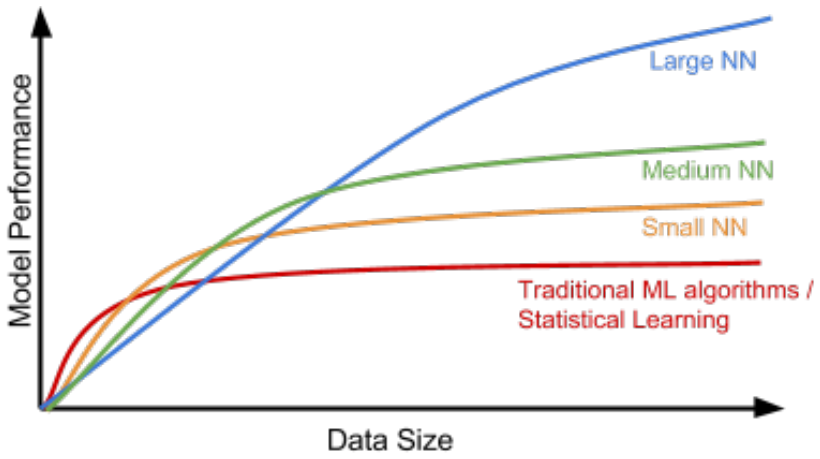


Figura 3.4: Relación entre el tamaño del conjunto de datos utilizado y el rendimiento obtenido con las diferentes técnicas. Se puede observar que las tradicionales funcionan mejor en conjuntos de datos reducidos, y según aumenta el tamaño son las redes neuronales las que ofrecen un mayor rendimiento. Fuente:[5]

De aquí surge la relación entre el tamaño de datos y el rendimiento del modelo [7]. Si la can-

tidad de datos es reducida, los algoritmos tradicionales o el aprendizaje estadístico hacen un gran trabajo. Pero cuando se aumenta esta cantidad de datos es cuando las redes neuronales aumentan su eficacia respecto a los demás tipos como podemos ver en Figura 3.4.

Estas estructuras básicas de la red neuronal dan lugar a diferentes tipos de estructuras, que se pueden clasificar de diferentes formas según el criterio utilizado:

***Número de capas*:**

- *Red neuronal monocapa*: Son las que únicamente tienen una capa de neuronas en las que se realiza el cómputo, es decir, solo tienen una capa intermedia además de las capas de entradas y salidas. Es el tipo de red más sencilla.
- *Red neuronal multicapa*: Extensión de las anteriores añadiendo un mayor número de capas intermedias entre la entrada y la salida.

***Tipo de conexión*:**

- *Red neuronal No recursiva*: En estas redes la propagación de la señal es únicamente en un sentido, no permitiendo retroalimentaciones. Estas redes no disponen de memoria.
- *Red neuronal recursiva*: Son un tipo de red neuronal que no toma únicamente como entrada el dato que les llega, sino que también utilizan lo que han percibido previamente. Esto quiere decir que la decisión tomada el paso de tiempo de una red recursiva $t-1$, afecta a la decisión que se alcanzará en el tiempo t. Es decir, toman el presente y el pasado como fuente de entrada para tomar decisiones.

***Grado de conexión*:**

- *Red neuronal totalmente conectada*: Cada una de las neuronas pertenecientes a una capa están conectadas con la siguiente capa y con la anterior en el caso de que fuese una red neuronal recursiva.
- *Red neuronal parcialmente conectada*: No existe una conexión total entre las neuronas de una capa con la siguiente.

3.3. Geometría solar

La radiación solar es un conjunto de radiaciones electromagnéticas emitidas por el sol, emitidas en un amplio espectro de frecuencias (luz visible, infrarrojo y ultravioleta), las cuales se producen en la fuente y son emitidas hacia fuera en todas las direcciones. Estas ondas no necesitan un medio material para propagarse, pueden atravesar el espacio interplanetario y llegar a la Tierra desde el Sol. Gran parte de la radiación que recibimos en nuestro planeta es detectable por el ojo humano, y constituye la luz visible (44 %), el resto son en gran medida infrarroja (49 %) y en una pequeña proporción ultravioleta (7 %).
Esta energía es el motor que mueve nuestro medioambiente, siendo la energía solar que llega a la

superficie terrestre 10.000 veces mayor que la energía consumida actualmente por toda la humanidad.

La radiación solar es la encargada de elevar la temperatura de los objetos y el suelo sin calentar el aire. Todos los cuerpos emiten radiación en función de su temperatura. Esta viene determinada por la ley de Stefan-Boltzmann, la cual determina que la energía emitida por un cuerpo negro por unidad de área y de tiempo es proporcional a la cuarta potencia de su temperatura absoluta.

$$E = \sigma * T_e^4 \tag{3.1}$$

donde T_e es la temperatura absoluta y σ es la constante de Stefan-Boltzmann $\sigma = 5{,}67x10^{-}8\frac{W}{m^2*K^4}$
Por cuerpo negro se entiende aquel que es capaz de absorber o emitir toda la radiación que incide sobre él.

Existen dos magnitudes para cuantificar la radiación solar, que corresponden a la potencia y la energía que se recibe por unidad de superficie:

***Irradiancia*:** describe la radiación que llega a la superficie terrestre. Es la potencia recibida por unidad de superficie, se expresa en W/m2.

***Irradiación*:** cantidad de irradiancia recibida en un lapso de tiempo. Se expresa en Wh/m2.

La radiación solar puede clasificarse en función de como las reciben los objetos:

***Radiación directa*:** Es la procedente del disco solar sin sufrir ningún tipo de alteración en su dirección. Se mide en los elementos que estén perpendiculares al sol.

***Radiación difusa*:** Es la que no proviene directamente del disco solar, sufre una modificación en su trayectoria original una vez llega a la atmósfera por diferentes motivos (densidad atmosférica, partículas, reemisiones de cuerpos...). Se mide sobre cualquier superficie eliminando/sombreando la componente directa.

***Radiación reflejada*:** Es la radiación reflejada por la superficie terrestre y objetos circundantes. Varía en función del coeficiente de reflexión de los distintos elementos reflectantes.

***Radiación global*:** Es la suma de las 3 radiaciones anteriores.

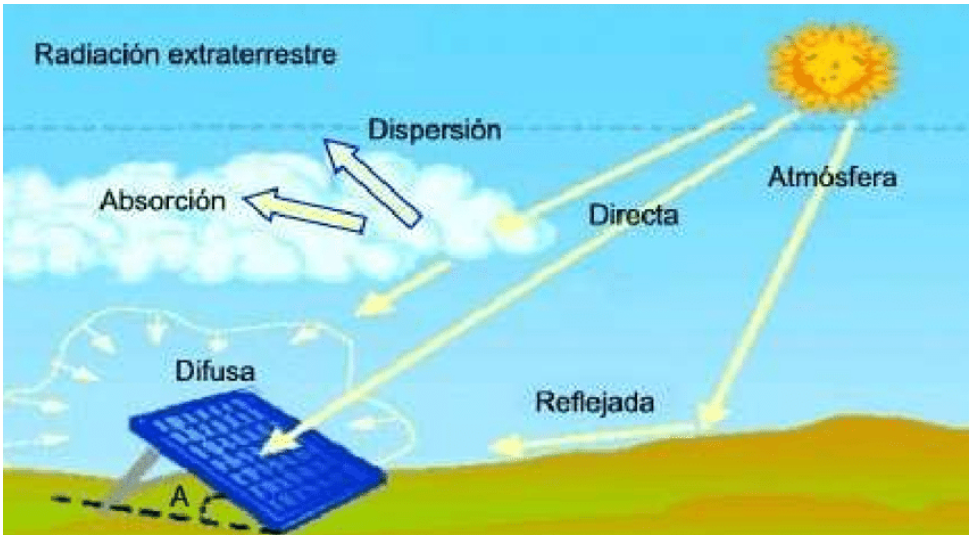


Figura 3.5: Esquema que muestra el comportamiento de los diferentes tipos de radiación. Fuente:[8]

La radiación no se recibe de forma constante en todo el planeta, sino que depende de la latitud, del ángulo de incidencia de la radiación solar y el número de horas de sol es un dato variable según las características climáticas del punto del planeta donde nos encontremos. En España los valores más elevados de radiación global anual se encuentran en la mitad sur de la península. No obstante, los valores máximos se encuentran en las zonas de menor nubosidad de Canarias, debido a su latitud tropical y el gran número de horas de sol.

3.3.1. Hardware para la lectura de la radiación

El instrumento que se utiliza para medir radiación global, difusa y reflejada se denomina piranómetro. Este sensor se encarga de medir la densidad de flujo de radiación solar (kilovatios por metro cuadrado) dentro de un rango de longitud de onde de 0.3μm a 3 μm

El espectro de radiación solar que llega a la tierra tiene una longitud de onda entre 0.3μm y 2.8μm. Para realizar la medición es necesario que la respuesta al haz de radiación varíe según el angulo de incidencia.

Existen principalmente tres tipos de piranómetros, que funcionan con dos tipos de tecnología diferente: termopila y semiconductores de silicio. La sensibilidad a la luz, conocida como respuesta espectral, depende del tipo de piranómetros. Esto significa que la respuesta será 1 cuando se encuentre de forma perpendicular y 0 cuando sea desde el horizonte.

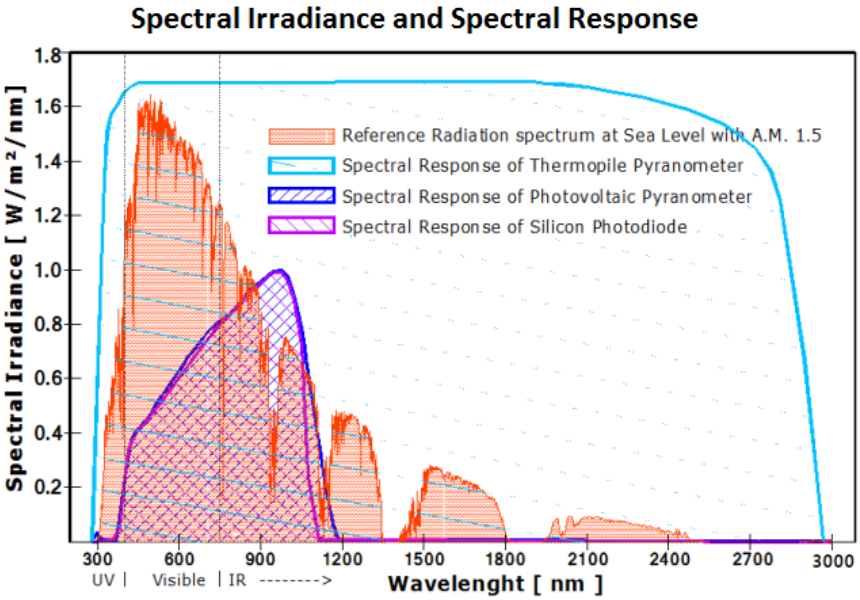


Figura 3.6: Gráfica que muestra las diferentes respuestas espectrales de cada tipo de piranómetro. Fuente:[9].

En la Figura 3.6 el espectro de irradiación representa la luz solar que alcanza la superficie terrestre a nivel del mar, a mediodía con una masa de aire de 1.5. En este espectro influyen factores como la latitud, la longitud, aerosoles o contaminación.

Los tres tipos de piranómetros son:

***Piranómetro térmico*:** Este tipo esta compuesto por una pila termoeléctrica alojada entre dos semiesferas de cristal. El principio físico utilizado es la medida de un termopar sobre el que incide la radiación a través de las dos cúpulas de vidrio. Estos piranómetros tienen la parte activa del sensor dividida en dos sectores: uno blanco y otro negro. La irradiación se calcula mediante la diferencia de temperatura entre la parte negra (que recibe la radiación solar) y la parte blanca.

***Piranómetro basado en fotodiodo*:** Está basado en fotodiodos y puede detectar el espectro solar entre 400nm y 900nm. Los fotodiodos convierten las frecuencias del espectro solar en corrientes gracias al efecto fotoelectrico. La corriente generada será proporcional a la irradiancia recibida.

***Piranómetro fotovoltaico*:** Es una derivación del piranómetro de fotodiodo. El sensor está formado por una célula fotovoltaica que funciona cortocircuitado. La corriente de salida será proporcional a la radiación solar incidente. Esté tipo de piranómetros son más sensibles a pequeñas variaciones, ya que no tienen la inercia térmica, como ocurre en los piranómetros térmicos.

[3]Fuente: http://www.kippzonen.es

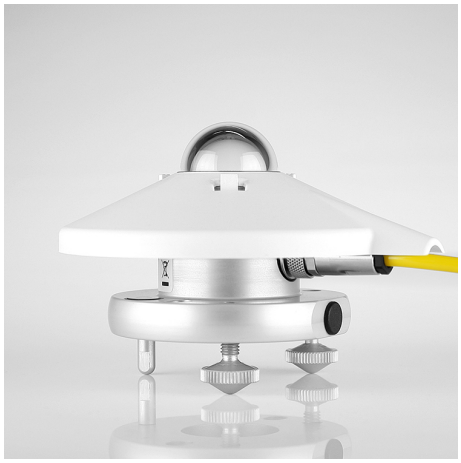


Figura 3.7: Ejemplo de un piranómetro térmico utilizado para medir la radiación solar.[3].

(a) Piranómetro basado en fotodiodo. (b) Piranómetro fotovoltaico.

Figura 3.8: Ejemplo de un piranómetros fotovoltaico y basado en fotodiodo utilizados para medir la radiación solar. Fuente:[9][5].

3.4. Modelo de cielo claro

Los modelos de cielo claro estiman la radiación solar global incidente en un lugar y momento determinado suponiendo un cielo despejado, es decir, sin ningún tipo de nubes que intercepten la

[5]Fuente: https://www.licor.com

llegada de la radiación solar. Los modelos más sencillos únicamente tienen en cuenta parámetros como: la altura solar, la latitud, la longitud, fecha, hora y la excentricidad de la órbita terrestre. Otros más avanzados tienen en cuenta factores como la concentración de componentes atmosféricos que puedan afectar, pero estos datos en ocasiones no están accesibles para ciertos lugares. Comúnmente estas dispersiones atmosféricas proceden de aerosoles, ozono y vapor de agua.
Estos modelos son frecuentemente utilizados y existe una gran cantidad de estudios para validar la eficacia de los diferentes modelos existentes. Son utilizados para normalizar los conjuntos de datos o conseguir que las series utilizadas se transformen en series estacionarias, ya que para modelar la radiación solar resulta muy interesante trabajar con series estacionarias.

3.5. Predicción en energía solar

Los modelos de predicción de radiación se suelen agrupar en tres tipos.
El primer tipo lo forman los modelos físicos basados en ecuaciones matemáticas que describen la física de la atmósfera[10], por ejemplo utilizando los vectores de movimiento de nubes obtenidos a partir de los vectores de movimiento de nubes[11][12].
En una línea diferente, los modelos estadísticos[13][14], establecen relaciones entre las observaciones pasadas y las futuras predicciones.
La tercera línea dentro de la que se encuentra nuestra propuesta, también emplea los datos históricos pero a través de modelos de aprendizaje automático tales como las redes neuronales[15][16][17].

La principal diferencia de nuestro trabajo con los ya mencionados es que estos se centran en predicciones normalmente de partir de 30 minutos[17], nosotros buscamos predicciones a muy corto plazo, de entre 1 y 10 minutos (nowcasting).

Capítulo 4

Metodología

En este capítulo se mostrarán las técnicas que se utilizarán para realizar la predicción y solventar los problemas tratados en el Capítulo 3, además de explicar el proceso de obtención de los datos.

4.1. Obtención de los datos

Los datos utilizados para realizar este estudio fueron ofrecidos por el CIEMAT, provenientes de la Plataforma Solar de Almería. Fueron obtenidos de las estaciones radiométricas ubicadas de la forma indicada en la Figura 4.1. De estas 19 estaciones fueron utilizadas 7, que se indican sus coordenadas exactas en el Cuadro 4.1 y en la Figura 4.2 se puede ver su ubicación en el mapa.

Figura 4.1: Ubicación exacta de las 19 estaciones que forman la forman la red radiométrica de la Plataforma Solar de Almería.

Las variables registradas en cada una de estas estaciones meteorológicas son irradiancia global

Figura 4.2: Ubicación exacta de las 7 estaciones de la Plataforma Solar de Almería utilizadas para la recolección de datos.

Nombre	Información Temporal	Latitud	Longitud	Altitud
BSRN (4)	UTC+0, Rec. freq. 60 s	37.092	-2.363	490.6
ARFRISOL (13)	UTC+0, Rec. freq. 60 s	37.094	-2.357	499.6
DISS (7)	UTC+0, Rec. freq. 5 s	37.098	-2.359	504.4
TSA (10)	UTC+1, Rec. freq. 1 s	37.093	-2.357	499.1
KONTAS (1)	UTC+1, Rec. freq. 1 s	37.095	-2.355	505.8
CESA1 (2)	UTC+1, Rec. freq. 60 s	37.095	-2.361	503.4
PSA-HP (3)	UTC+1, Rec. freq. 10 s	37.091	-2.358	500.0

Cuadro 4.1: Información de las estaciones utilizadas en el estudio.

horizontal, difusa, directa normal, temperatura, humedad relativa, velocidad y dirección de viento. En este estudio únicamente se utilizará la radiación global medida con piranómetros térmicos de la marca Kipp & Zonnen. Los datos una vez son leídos por los piranómetros son enviados a un servidor central mediante FTP en el cual se realizará el almacenamiento de los mismos.

Una vez recibidos los datos, se generan los ficheros para cada una de las estaciones, añadiendo la columna *timestamp* para llevar un registro temporal de cuando se realizaron las mediciones. Cada uno de estos ficheros almacena la información de un mes para una estación.

Como se puede observar en el Cuadro 4.1 algunas de estas estaciones ofrecían diferentes intervalos de muestreo, por tanto para poder trabajar con todos los datos se utilizó la frecuencia mayor, es decir, 60 segundos. Por tanto, la información que se utilizará en este estudio consiste en los datos que nos envían cada una de las estaciones meteorológicas, con una frecuencia de un minuto, que se almacenaran en ficheros con una extensión de un mes.

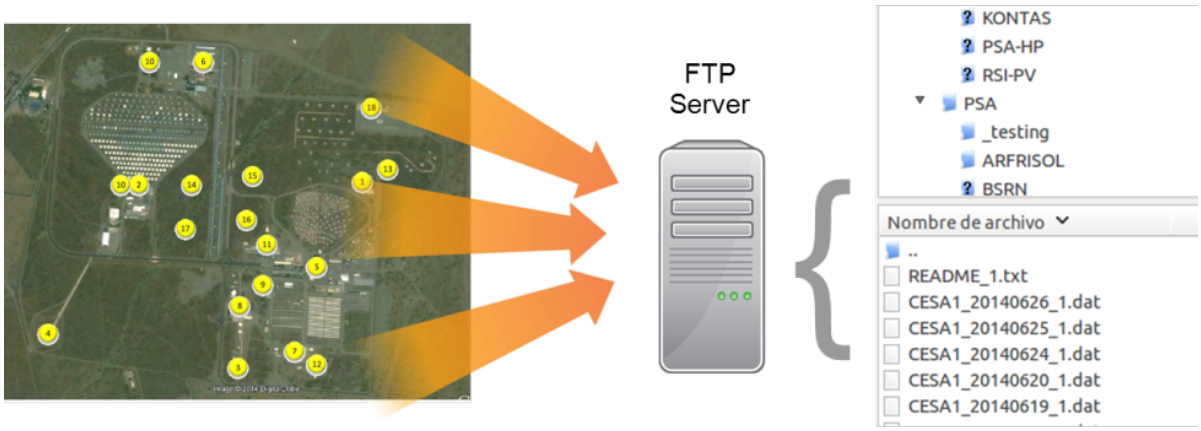


Figura 4.3: Esquema que indica el procedimiento que se utilizó para la recepción de los datos en las diferentes estaciones.

4.2. Fases del trabajo

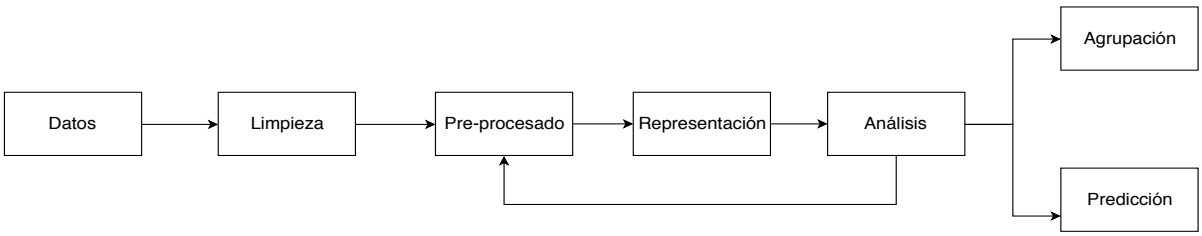


Figura 4.4: Diagrama que muestra la metodología utilizada para realizar la predicción.

Tras recibir los datos, la primera fase consistió en un pre-procesado de los mismos, en el cual, se realizaron comprobaciones para eliminar entradas ausentes o incorrectas. Tras esto se realizó una normalización y se filtró el número total de datos. Una vez realizados estos primeros ajustes, se realizaron representaciones para poder explorar los datos y encontrar posibles parámetros que modificar. Una vez terminada la fase de preparación de los datos, se realizó un agrupamiento de los datos para poder realizar una primera clasificación. A continuación se aplicó a los datos diferentes algoritmos de análisis de datos para poder realizar la predicción.

Tras esto se realizó una comparativa de los diferentes algoritmos utilizados. La métrica utilizada para realizar esta tarea es la raíz del error cuadrático medio (Root Mean Square Error, RMSE, por sus siglas en inglés). El RMSE es la desviación estándar de los residuos (errores de predicción). Se utiliza para medir el nivel de error que existe entre dos conjuntos de datos, es decir entre el

valor predicho y el valor conocido.

$$RMSE = \sqrt{\Sigma_{t=1}^{T}\Big(\frac{\hat{y}_t - y_t}{T}\Big)^2} \tag{4.1}$$

donde $\hat{y}_t$ son los valores de la predicción, y_t es la variable de referencia y T es el número de veces que se han realizado las observaciones.

4.3. Algoritmos de predicción

En primer lugar se realizó la comprobación para saber si el conjunto de datos utilizado sigue una serie temporal estacionaria. Esta serie temporal tiene la peculiaridad de que sus propiedades no dependen del momento en que se observa, siendo más difíciles de predecir, siendo por tanto conveniente realizar una transformación para así evitar la estacionalidad.
Para comprobarlo se ha utilizado el método de aumentado de Dickey–Fuller (ADF)[18], obteniendo como resultado que nuestro conjunto de datos sigue una serie temporal estacionaria, tanto en el conjunto de datos de forma anual como de forma diaria.
Para evitar la estacionalidad, se suele recurrir a la utilización de las diferencias, es decir, se reemplaza cada valor por el incremento, positivo o negativo, con respecto al dato anterior. Con esta transformación se pierde un dato por día (el primero de la mañana, donde la diferencia no tiene sentido), pero en nuestro caso esto carece de importancia dado el elevado volumen de datos del que disponemos.
Tras esta transformación se puede comprobar que ahora se obtienen datos estacionarios, de forma más clara en los datos diarios, y por muy escaso margen en el total anual. Esto nos lleva a decidir que emplearemos datos diarios para nuestros modelos.

4.3.1. Métodos directos

Vamos a utilizar dos métodos sencillos de predicción referenciados habitualmente en la bibliografía.

***Predicción naïve*:** La predicción es simplemente el último valor conocido. Es un método que tiene un coste muy eficiente y produce valores que se suelen utilizar como punto de referencia frente a métodos más complejos. Solo puede ser utilizado en series temporales.

$$\hat{y}_{T+h|T} = y_T \tag{4.2}$$

donde y_T es el dato anterior.

***Media*:** Se propone como predicción la media de los datos históricos conocidos.

$$\hat{y}_{T+h|T} = \bar{y} = (y_1 + ... + y_T)/T \tag{4.3}$$

donde $y_1, ..., y_T$ son los datos pasados.

La (4.3) hace referencia a series temporales ya que son de ese tipo los datos utilizados en el estudio, pero este tipo de predicción no es únicamente para este tipo de series, puede ser utilizada también para otros tipos.

A menudo, sobre todo en datos con alta estacionalidad, estos métodos sencillos son los más efectivos a corto plazo. En nuestro caso, donde consideramos horizontes de muy pocos minutos, veremos que el método naïve resulta muy difícil de superar.

4.3.2. Auto Regressive Integrated Moving Average

ARIMA (Auto Regressive Integrated Moving Average) [18] [19] es un tipo particular de modelo predictivo muy utilizado para series temporales, en los cuales se tiene en cuenta la relación existente entre los datos, es decir, cada registro de un instante preciso está generado en función de los valores anteriores. Este método permite especificar un valor en una función lineal de valores anteriores y errores provocados por el azar. Sus tres componentes principales son auto regresivo, integrado y medias móviles, los cuales vienen representados respectivamente por q, d, p.
El parámetro q nos muestra el número de errores observados en las anteriores predicciones. La d nos indica la cantidad de diferencias que deben realizarse para convertir la serie en estacionaria. Por último la p nos indica el número de registros anteriores que se utilizarán para realizar la predicción del siguiente valor.
La metodología de ARIMA se puede dividir en cuatro fases:

- La primera consiste en determinar el posible modelo de ARIMA que sigue la serie. Para ello se decide que transformación aplicar para convertir la serie en no estacionaria y tras esto determinar el modelo.
- La segunda, una vez seleccionado el posible modelo, consiste en fijar los órdenes de p y q que se utilizarán para dicho modelo.
- En la tercera fase se comprueba que los residuos siguen un proceso de ruido blanco, es decir, su media es cero y su varianza es constante para diferentes valores. Si los residuos siguen algún tipo de estructura se deberá repetir el proceso de selección de modelo.
- La cuarta fase consiste en realizar la predicción con el modelo elegido.

4.3.3. Redes long short-term memory

Dentro de las redes neuronales existe un tipo llamado recurrent neural network (RNN) donde las conexiones entre nodos forman un grafo dirigido. Las neuronas reciben como entrada la salida de la capa anterior y propagan su salida a la siguiente, permitiendo así cierta memoria y, por tanto, un sentido temporal. Debido a su capacidad para utilizar su estado interno para procesar secuencias de entrada, las RNN son muy utilizadas para la identificación y clasificación de patrones secuenciales con diferentes probabilidades de repetirse en el tiempo.
Las redes long short-term memory (LSTM) son una implementación concreta de las RNN, remplazando las tradicionales neuronas por una estructura más compleja llamada de memoria LSTM.

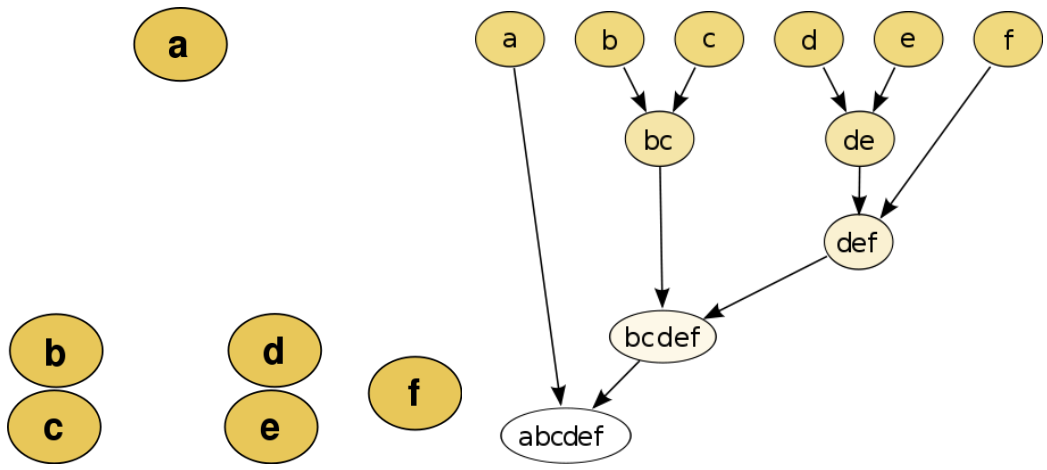
Estas celdas disponen de tres puertas para controlar el estado de la celda. La puerta del olvido se encarga de decidir qué información enviaremos del estado de la célula. La de la entrada controla cuando la información nueva puede entrar en la memoria. Esta tiene dos capas, una primera basada en una función sigmoide que decide qué valores actualizaremos, y después una capa con una tangente hiperbólica que se encarga de crear un nuevo array de valores candidatos. Por último, la puerta de la salida indica qué información de la contenida se utiliza para el resultado.

4.4. Algoritmos de agrupación

El agrupamiento o *clustering* es una técnica de minería de datos dentro de la disciplina de la inteligencia artificial, que consiste en agrupar un conjunto de objetos de tal manera que los objetos de un mismo grupo (*cluster*) sean, de cierto modo, más similares entre sí que a los otros elementos del resto de grupos. Es decir, se busca similitud intra-cluster alta y similitud inter-cluster baja.

Existen dos grandes técnicas para el agrupamiento:

***Agrupamiento jerárquico*:** tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos. Esta técnica genera un árbol que dependiendo de una matriz de proximidad, indica la relación entre los objetos. Cada hoja es un elemento y el nodo raíz representa el conjunto total de datos. Para obtener los diferentes clusters se debe partir el árbol a diferentes niveles. Si es de la raíz a las hojas es divisivo, y si por el contrario es de las hojas a la raíz se llama aglomerativo.



(a) Grupo de datos original. (b) Árbol generado mediante el agrupamiento jerárquico.

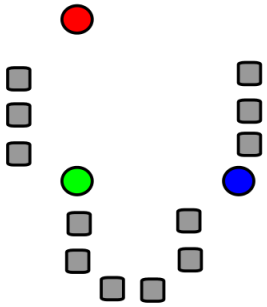
Figura 4.5: Grupo original de datos y árbol generado mediante el agrupamiento jerárquico.Fuente:[9].

***Agrupamiento no jerárquico*:** en esta técnica, al contrario que en los agrupamientos jerárquicos, se debe saber el número de grupos de antemano, y posteriormente cada valor se asigna a los

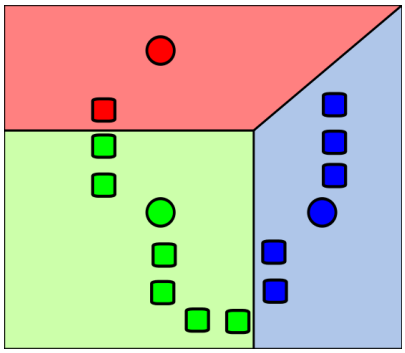
grupos en función de su cercanía. Un ejemplo de implementación concreta de un algoritmo es *k-means*.

4.4.1. *k-means*

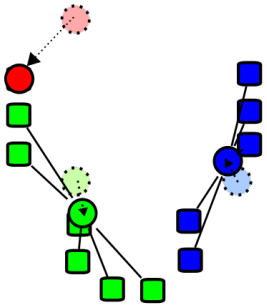
k-means es un algoritmo de clasificación no supervisada y no jerárquico que agrupa objetos según sus características. Su objetivo es dividir n observaciones en k clústeres. La forma que tiene de agrupar es minimizando la suma de las distancias entre cada objeto y el centro del clúster. La medida estándar es la distancia cuadrática.



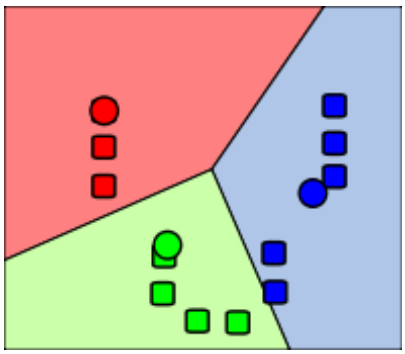
(a) Elección de K y posicionamiento de los centros



(b) Primera asociación de los elementos a los centros



(c) Reposicionamiento de los centros respecto a los elementos que tienen asignados



(d) Repetición de los pasos anteriores hasta finalizar la clasificación.

Figura 4.6: Las imágenes muestran los diferentes pasos que se ejecutan para poder realizar la clasificación de los elementos. Fuente:[9].

Este algoritmo usa una técnica de refinamiento iterativo que consta de las siguientes partes:

1. En primer lugar se deben elegir el número de grupos, K, que vamos a utilizar. Una vez escogidos se establecen los centros de cada uno. Para esta primera posición de los centros se pueden elegir, por ejemplo, de forma aleatoria.
2. Cada uno de los elementos del conjunto se asignarán al centro más próximo, es decir, que la distancia cuadrática sea mínima.

3. Tras asignar todos los objetos a un centro, se deberán actualizar cada uno de los centros de los grupos a la posición promedio de todos los integrantes de ese grupo.
4. Por ultimo, se repetirán los pasos de asignar los elementos al centro y de reposicionar los centros hasta que estos se muevan por debajo de un umbral.

Capítulo 5

Procesado de datos

Una vez obtenidos los datos, lo primero que se debe realizar es una tarea de procesado y filtrado. Los datos estaban separados en diferentes archivos, los cuales almacenaban la información de una estación y un meses concreto. Por tanto, para facilitar la predicción, se unificaron todos los datos en un archivo cuya extensión era mensual, siendo cada una de las columnas una estación.

```
TIMESTAMP,AFRISOL,CESA1,DISS,PSA,TSA,KONTAS,BSRN
2015-03-01 00:00:00,-1.453,-2.902,-7.6,-3.322,-1.057,5,-4.7
2015-03-01 00:01:00,-1.453,-2.902,-6.9,-3.206,-1.107,5,-4.9
2015-03-01 00:02:00,-1.453,-2.902,-7.6,-3.129,-1.157,5,-5
2015-03-01 00:03:00,-1.453,-2.902,-7.6,-3.129,-0.956,5,-5.1
2015-03-01 00:04:00,-1.453,-2.902,-6.9,-3.129,-1.006,5,-4.9
2015-03-01 00:05:00,-1.635,-2.902,-6.3,-3.129,-1.006,5,-4.8
2015-03-01 00:06:00,-1.453,-2.902,-6,-3.129,-1.057,5,-4.7
2015-03-01 00:07:00,-1.453,-2.902,-6.9,-3.129,-1.057,5,-4.6
2015-03-01 00:08:00,-1.453,-2.902,-6.3,-3.129,-1.157,5,-4.6
2015-03-01 00:09:00,-1.635,-2.902,-6.9,-3.129,-1.157,5,-4.6
2015-03-01 00:10:00,-1.817,-2.902,-6,-3.129,-1.107,5,-4.7
2015-03-01 00:11:00.-1.817.-2.902.-5.6.-3.129.-1.107.5.-4.9
```

Figura 5.1: Ejemplo de los datos originales recibidos por las estaciones en marzo de 2015 tras realizar la unión de todas las estaciones en un único archivo.

Estos archivos finales consistían en 8 columnas, la primera era la fecha y las otras 7 eran la radiación solar recibida por cada uno de los sensores en la fecha indicada. También se efectuó un primer filtrado de los datos convirtiendo a 0 todos los valores negativos, ya que no se puede dar una radiación negativa. Estos valores son producidos debido al margen de error en la lectura por parte de los sensores en situaciones donde la radiación es baja.

5.1. Datos ausentes

A continuación se realizó un primer procesado buscando valores de los sensores ausentes, es decir, minutos en los que, por alguna razón, no se hubieran recibido información desde dicho

sensor. Aunque escasos en el conjunto de datos, la ausencia de datos puede producir alteraciones o incluso imposibilitar la predicción. Ante estas situaciones se suelen plantear dos formas de actuar. La primera consiste en reemplazar estos valores ausentes por otro valor, comúnmente la media de los otros valores, en este caso la media del resto de estaciones en el minuto dado. La otra forma consiste en eliminar los registros que se ven afectados por esta situación, en este caso consistiría en eliminar la fila completa del minuto en el que ocurrió esta situación. La opción utilizada en este trabajo fue reemplazar los valores ausentes por la media del resto de estaciones en el minuto dado, evitando así reducir el número de datos que se utilizarán para la predicción.

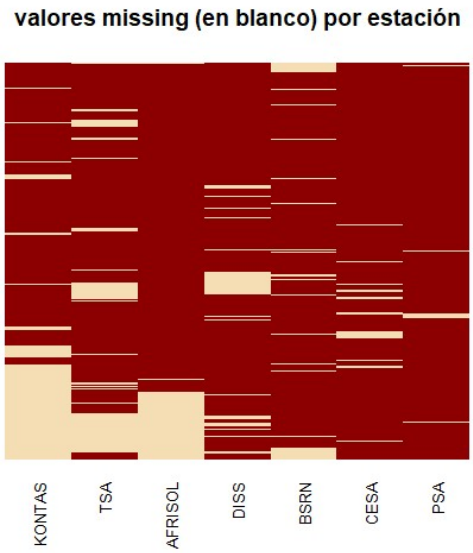


Figura 5.2: Gráfica del número de valores ausentes en el conjunto de datos utilizado para cada una de las estaciones meteorológicas.

5.2. Ajuste de la franja horaria de predicción

La siguiente tarea a realizar fue un ajuste de la franja horaria utilizada para la predicción, ya que no es interesante utilizar las horas de noche debido a que no existe radiación solar y por tanto no se puede obtener energía durante esa franja. Por tanto, para minimizar el efecto se ha optado por establecer la misma franja horaria para todos los meses de año, empezando a las 9:00 horas y terminando a las 16:00 horas. Esto reduce el conjunto de datos a 7 horas al día en lugar de las 24 horas que teníamos en un principio.

5.3. Ajuste del retraso de las estaciones

Una vez realizado el primer procesamiento, se procedió a representar los datos para poder obtener más información de los datos de forma visual. Esta representación se hizo plasmando en una

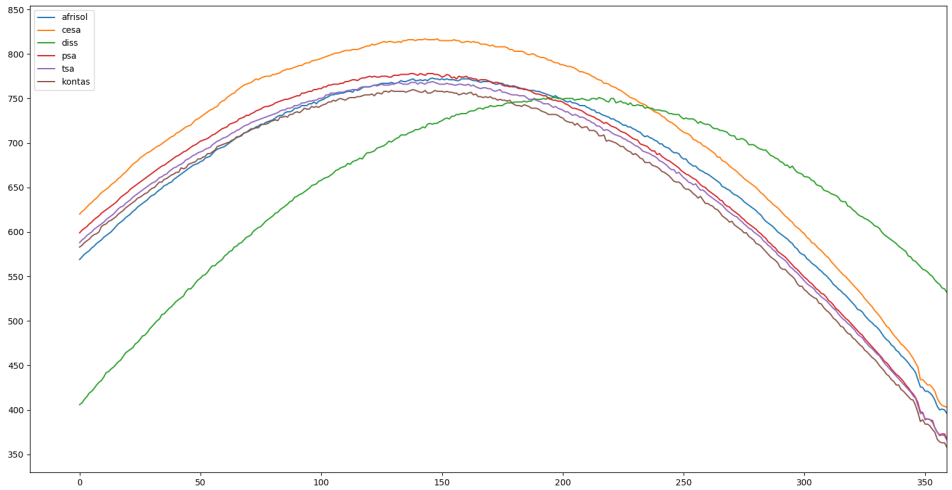


Figura 5.3: Gráfica que representa la radiación solar de todas las estaciones entre las 10:horas y las 14:00 horas. El eje X representa la radiación solar y el eje Y el timestamp. Se puede observar como la línea naranja y verde se encuentran desplazadas respecto a las demás.

gráfica la evolución de la radiación solar global para cada día. Además se realizaron vídeos para poder comprobar de forma continua como evolucionaban durante cada día. Estos vídeos eran uno por cada mes, y en ellos se representaba en escala de grises, sobre un fondo de la ubicación real de las estaciones, las diferentes radiaciones obtenidas por los diferentes sensores. Para poder ubicar cada uno de los sensores se realizó una matriz de 10x10 en la cual 7 casillas estaban ocupadas por los sensores. El resto se calcularon mediante interpolación. Cuanto mayor era la radiación, implicaba que el cielo estaba más despejado, por tanto el color era más cercano a blanco, en cambio si era un valor alto sería cercano a negro. Gracias a estas representaciones se pudo observar como algunas estaciones sufrían retrasos respecto a las demás. Tenían el mismo comportamiento que el resto de sensores, pero ocurría unos minutos más tarde.

Para comprobar esto se buscaron días despejados del mes donde se encontró el comportamiento anómalo, y se representaron todos los sensores entre las 10:00 y las 14:00. Esto se debe a que entorno a las 12:00 horas se encuentra el punto máximo de radiación, por tanto al representarlos todos se puede observar en que punto registra cada uno el valor más alto de radiación global. Esto concluyó en que tanto la estación *DISS* como *KONTAS* sufrían retrasos en sus valores.

Para la estación *KONTAS* se utilizó un factor de corrección c, que nos indicaba el número de minutos. Este parámetro c nos indicaba el número de minutos que estaba retrasada la estación.

Para poder fijar el valor se tomo k (la estación *KONTAS*) y r otra cualquiera.

$$error = \sum_{t} ||v(k, t + c) - v(r, t)|| \tag{5.1}$$

Donde c indica el número de minutos de retraso, t el minuto que corresponde y $v(k, t)$ y $v(r, t)$ es la radiación obtenida en el momento t por las estaciones k y r. El parámetro c es el utilizado para minimizar la función.

Una vez realizado este proceso iterando con diversos valores de c se consiguió el error mínimo para un valor de 16. Esto quiere decir que la estación KONTAS tenía 16 minutos de retraso con el resto de estaciones.

Para la estación *DISS* se hizo algo similar, pero utilizando también la siguiente fórmula:

$$error = \sum_{t} ||v(k, t) * c - v(r, t)|| \tag{5.2}$$

Donde c indica el factor por el que multiplicar la radiación obtenida, t el minuto que corresponde y $v(k, t)$ y $v(r, t)$ es la radiación obtenida en el momento t por las estaciones k y r. El parámetro c es el utilizado para minimizar la función.

Se aplicaron ambas fórmulas ya que *DISS* daba, al igual que *KONTAS*, valores con retraso pero además los valores eran mayores. Tras intentar minimizar el error se pudo observar que la estación *DISS* presentaba retrasos irregulares difíciles de predecir. Debido a esto nos vimos en la obligación de eliminar los datos de esta estación del conjunto utilizado para realizar la predicción, ya que podría afectar al resultado final su extraño comportamiento.

5.4. Modelo de cielo claro

Para optimizar la eficacia de los algoritmos de predicción es conveniente normalizar los valores en un rango definido. Es este caso se optó por aplicar un modelo de cielo claro para realizar el proceso de normalización. Como se explicó en el Capítulo 3, el modelo de cielo claro permitía estimar el valor de la radiación solar global en un cielo sin nubes, utilizando parámetros como el ángulo de elevación, altitud del lugar, y de otras condiciones atmosféricas.
Una vez que sepamos cuál es la radiación R esperada en condiciones de cielo despejado en ese momento y en ese lugar, podemos dividir la radiación obtenida por el sensor r por este valor. El valor r/R nos dará la proporción de radiación obtenido con respecto al máximo posible, que debe ser un valor entre 0 y 1.

En primer lugar se utilizó un modelo de cielo claro que ofrecía una librería de Python para realizar este proceso. Está librería era Pysolar, la cual permitía obtener el valor estimado de cielo despejado para un lugar y un momento determinado. Para cada día se calculó el valor dado por el modelo y se utilizó para dividir el número dado por el sensor. Los resultados obtenidos no fueron satisfactorios, ya que la normalización no se encontraba entre 0 y 1 ni seguía el comportamiento habitual del aumento y decremento durante el día de la radiación solar.

Por tanto se procedió a utilizar un modelo ofrecido por una web, que tras testearlo, se comprobó que ofrecía resultados similares a los esperados. La forma de obtener los valores de este

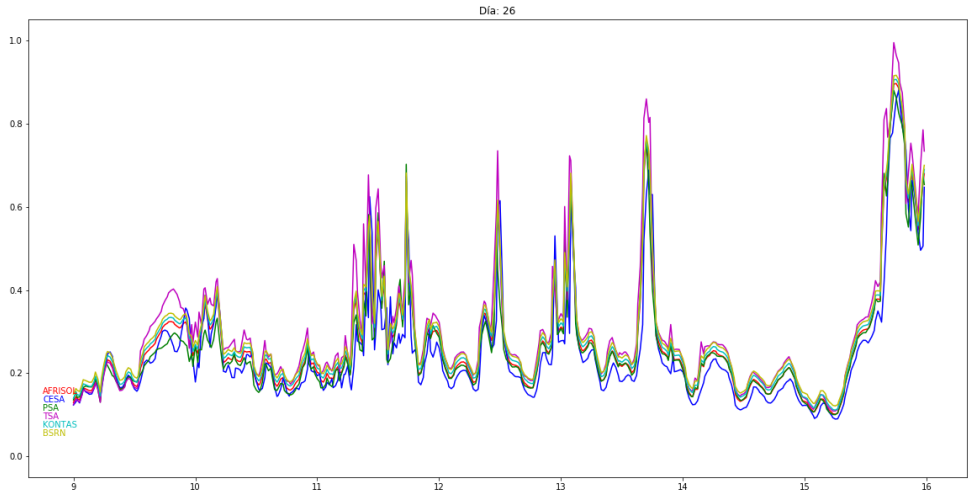


Figura 5.4: Gráfica de radiación del día 25 de enero del 2015. Este es un ejemplo de un día nuboso. El eje Y muestra el cociente entre la radiación global y la radiación de cielo despejado y el eje X indica la hora del día.

modelo fue mediante *web scraping*, rellenando mediante un script de forma automática el formulario, enviándolo y leyendo del campo correspondiente la respuesta obtenida.

En este caso una vez aplicado el modelo se consiguieron valores normalizados entre 0 y 1 que seguían el comportamiento habitual de la radiación solar.

Las Figuras 5.4 y 5.5 muestran un ejemplo de día nuboso y día despejado respectivamente. Para facilitar la visualización se han modificado ligeramente las coordenadas Y de cada estación, que de otra manera quedarían superpuestas. El eje Y muestra el cociente entre la radiación global y la radiación dada por el modelo de cielo claro y el eje X indica la hora del día en la que se produjo.

La 5.4 es del día 25 de enero del 2015. Podemos comprobar como ha sufrido variaciones durante todo el día, eso nos indica que ha sido un día principalmente nuboso, pero con ciertos momentos despejado.

La 5.5 es del día 26 de agosto del 2015 y es un claro ejemplo de un día despejado, durante todo el día se ha mantenido constante el valor del cociente siendo este de entorno a 0.8.

5.5. Preparado para predicción

Por último se aplicó una última fase de corrección de errores provocados por el proceso de normalizar.

La figura 5.6 es del día 11 de octubre del 2015 y muestra un ejemplo de estos últimos errores

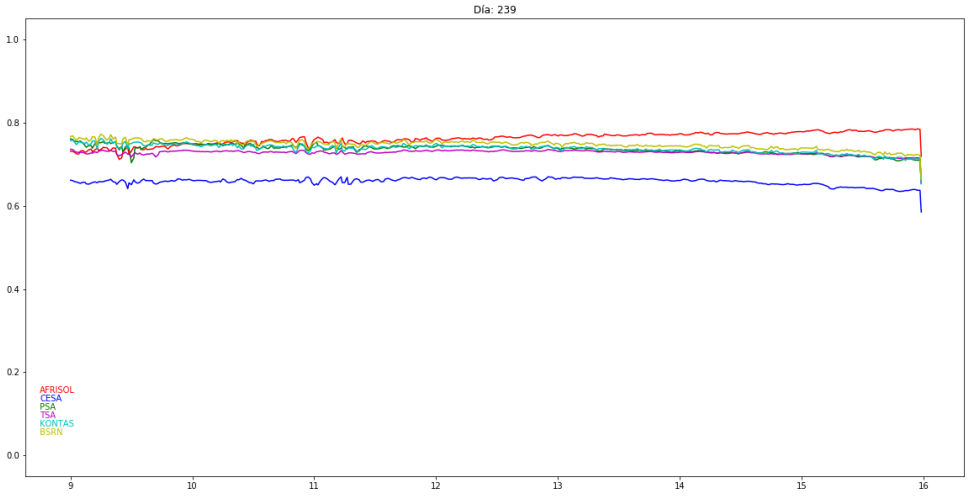


Figura 5.5: Gráfica de radiación del día 26 de agosto del 2015. Este es un ejemplo de un día despejado. El eje Y muestra el cociente entre la radiación global y la radiación de cielo despejado y el eje X indica la hora del día.

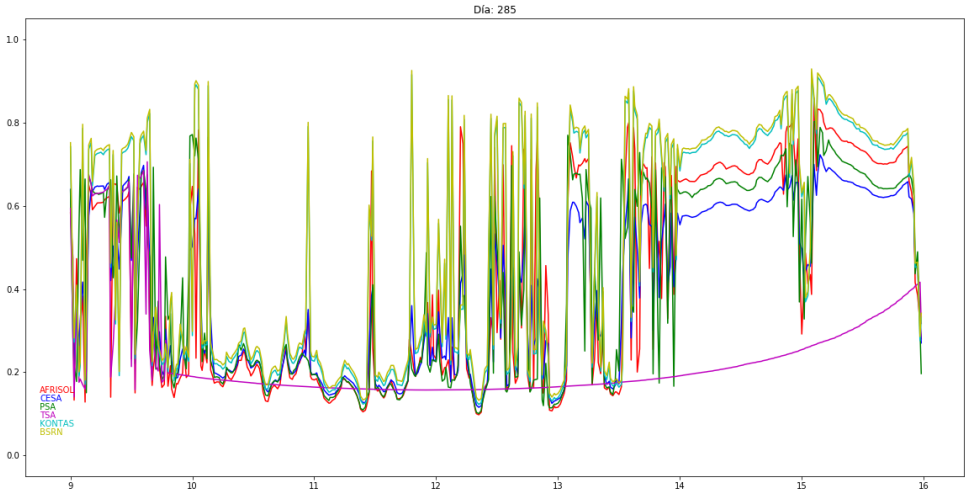


Figura 5.6: Gráfica de radiación del día 11 de octubre del 2015. Este es un ejemplo de un día en el que la TSA sufre un error. El eje Y muestra el cociente entre la radiación global y la radiación de cielo despejado y el eje X indica la hora del día.

que corrigieron. Se puede observar como el comportamiento que tiene la estación TSA durante todo el día es completamente diferente al resto de estaciones, por tanto se puede deducir que ha sufrido alguna anomalía y debe ser corregida.

Ya una vez normalizados los datos y filtrados correctamente, se procedió a unificar todos los ficheros en uno, el cual tenía la información de todo el año, entre las 9:00 y las 16:00 horas con una resolución temporal de 1 minuto.

Capítulo 6

Resultados

En el Capítulo 4 se explicó cómo se realizó la obtención de los datos y cuáles fueron los algoritmos que se utilizaron para realizar la predicción. Tras ello, en el Capítulo 5, se mostró cuál fue el procesado que se realizó una vez obtenidos los datos para poder utilizarlos en el estudio. A continuación, en este capitulo se mostrará cuáles han sido los resultados obtenidos en la predicción y en el proceso de segmentación.

6.1. Búsqueda de cielo nuboso mediante segmentación

Un primer análisis de interés consiste en comprobar si los datos admiten alguna segmentación, que permita agrupar mediciones similares y de esta forma determinar un número pequeño de posibles escenarios a analizar.

6.1.1. Número idóneo de clústers

Uno de los métodos más comunes para este propósito es el método propuesto por T. Caliński y J. Harabasz[20]. En este método, los valores a segmentar se consideran puntos dentro de un espacio n-dimensional euclídeo, para los que hay que encontrar una partición óptima. En nuestro caso tendríamos puntos de un espacio de 6 dimensiones, donde cada dimensión corresponderá al valor de una de las estaciones.

La definición de partición óptima en este método se corresponde a aquella partición que permitan minimizar la suma de los cuadrados de las distancias a los centros de cada segmento.

Los detalles de este algoritmo están más allá del ámbito de este trabajo. En todo caso se encuentra implementado en forma de librerías de los lenguajes Python (librería sklearn), y R (librería vegan). En ambos casos da como número óptimo el valor k=2.

6.1.2. Proceso de segmentación, e interpretación de los resultados

Los centros de cada segmento no se obtienen directamente con este método, pero una vez obtenido que el K óptimo es 2, se podrá emplear algún método de clasificación como k-means para obtener esos valores, así como una clasificación del conjunto de datos utilizando 2 clúster. Los resultados obtenidos se podrán comprobar en el Cuadro 6.1.

Nombre	Clúster1	Clúster2
ARFRISOL	0.259	0.682
CESA	0.264	0.679
PSA	0.269	0.685
TSA	0.268	0.690
KONTAS	0.277	0.703
BSRN	0.286	0.703

Cuadro 6.1: Centro de los clúster para cada estación. El Clúster1 indica cielo nuboso y el Clúster2 indica cielo despejado.

La primera columna indica el nombre de cada una de las estaciones, mientras que las otras dos indican el valor central para cada uno de los dos clústers. La interpretación más natural es que el primer grupo corresponde a una medición de nubes, mientras que el segundo se correspondería a un estándar de cielo despejado. Un vistazo al histograma de frecuencias para cada una de las estaciones parece confirmar esta interpretación. En particular, la Figura 6.1 contiene el histograma de la estación PSA-HP.

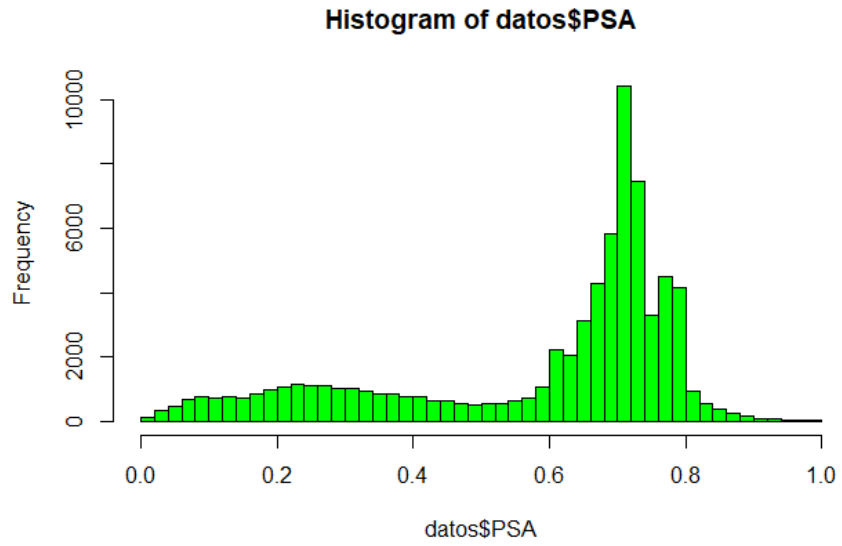


Figura 6.1: Histograma de frecuencias para las distintas radiaciones registradas en la estación PSA-HP

En efecto, la Figura 6.1 parece mostrar una combinación de dos distribuciones. Los centros del clúster 1 se corresponden con el centro de la distribución situada a la izquierda, y por tanto de menor radiación. Esto indicaría que se trata de los valores que corresponden a un cielo con nubes. En cambio, la segunda distribución, más apuntada, corresponde al cielo despejado, que como se aprecia en el histograma es más frecuente en las estaciones estudiadas.

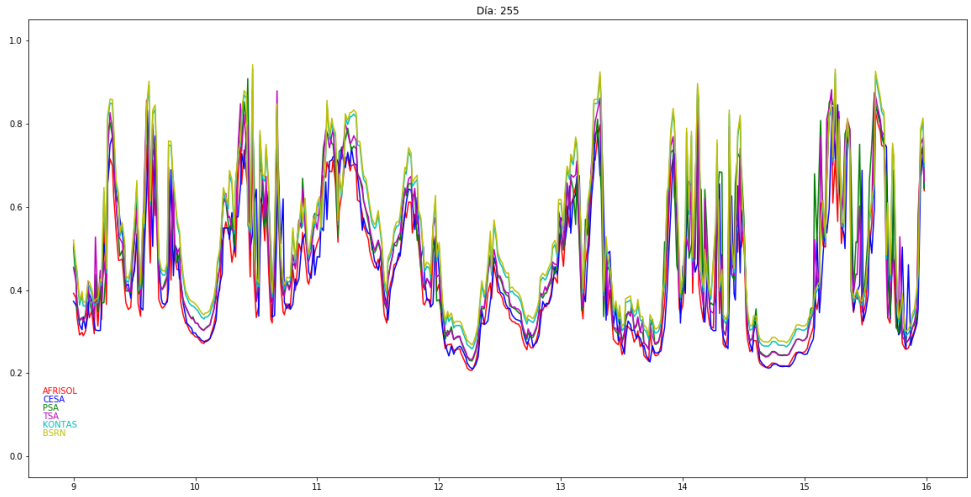


Figura 6.2: Gráfica de radiación del día 11 de septiembre del 2015. El eje Y muestra el cociente entre la radiación global y la radiación de cielo despejado y el eje X indica la hora del día.

Observando la Figura 6.2 podemos comprobar como durante el día 11 de septiembre de 2015 hubo variaciones en la radiación obtenida. Teniendo en cuenta los valores de los centros de los clúster de la Cuadro 6.1, se puede comprobar que en las zonas más bajas los valores son entorno 0.2 y 0.3, siendo esto cielo nuboso. En cambio los picos más altos de radiación se encuentran aproximadamente entre 0.8 y 0.9, siendo momentos donde el cielo está despejado.

6.2. Resultados predicción

El siguiente paso a realizar fue aplicar los procesos y los algoritmos explicados en el Capítulo 4, determinando los parámetros de cada algoritmo para realizar la predicción.
El método de ARIMA se ha utilizado la biblioteca de R ”forecast”, y en concreto el método ”auto.arima”, obteniendo unos valores de $p = 5$, $d = 1$ y $q = 0$.

La red neuronal LSTM se ha programado mediante la librería ”Python Keras” (https://keras.io/) funcionando sobre la librería ”TensorFlow” (https://github.com/tensorflow/tensorflow). ”Keras” admite numerosas funciones de coste (parámetro loss), entre las que hemos elegido la función *mean_squared_error*, dado que emplearemos la raíz del error cuadrático medio (RMSE) para determinar el error de cada método.

En este caso, se parte de una capa de entrada que recibe una única señal, y una única capa oculta, una capa de tipo LSTM. Se ha encontrado que el número de neuronas adecuado depende del horizonte considerado. En particular, en los experimentos, se ha empleado una red LSTM de 4

neuronas monocapa.

Las entradas que recibía esta red LSTM son los datos de una estación meteorológica, por tanto cada uno de las pruebas realizadas fue repetida en este caso 6 veces, una vez por cada una de ellas.

f (min)	LSTM (RMSE)	AVG (RMSE)	Naïve (RMSE)	ARIMA (RMSE)
1	0.054	0.355	0.382	0.383
2	0.070	0.522	0.142	0.081
3	0.093	0.214	0.106	0.118
4	0.11	0.409	0.115	0.126
5	0.12	0.362	0.123	0.139
10	0.15	0.370	0.144	0.175
20	0.21	0.287	0.195	0.214

Cuadro 6.2: RMSE resultante de la predicción de los valores de las estaciones utilizando LSTM, el método de la media, naïve y ARIMA con un horizonte de predicción f.

La primera columna de los Cuadros 6.2 y 6.3 indica los distintos horizontes temporales expresados en minutos utilizados para realizar las predicciones. El resto de columnas corresponden a las predicciones realizadas con los cada uno de los algoritmos de predicción. En el caso del Cuadro 6.2 muestra el valor de RMSE obtenido y para el Cuadro 6.3 la segunda columna indica el RMSE obtenido con la red neuronal basada en LSTM y el resto de las columnas muestra el factor multiplicador de diferencia entre el RMSE con cada uno de los métodos y el obtenido para la red LSTM.

f (min)	LSTM (RMSE)	AVG (factor inc)	Naïve (factor inc)	ARIMA (factor inc)
1	0.054	6.59	7.08	7.10
2	0.070	7.46	2.03	1.17
3	0.093	2.31	1.14	1.27
4	0.11	3.72	1.05	1.15
5	0.12	3.02	1.03	1.16
10	0.15	2.47	0.96	1.17
20	0.21	1.37	0.93	1.02

Cuadro 6.3: RMSE de la predicción LSTM modificada, junto con el factor multiplicador de incremento del RMSE del resto de métodos con respecto a LSTM modificado con el horizonte temporal f.

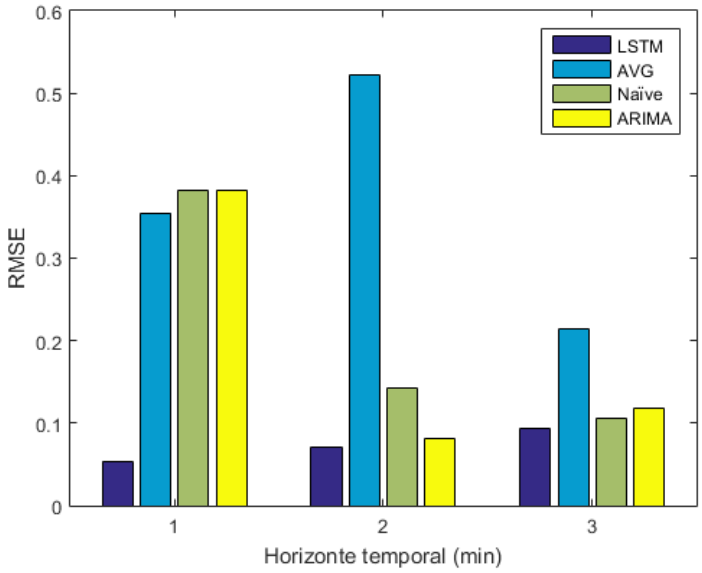


Figura 6.3: Gráfica comparativa entre los valores obtenidos de RMSE para los diferentes algoritmos de predicción.

El método de la media (AVG) ha obtenido los mayores valores de RMSE en los horizontes temporales de 1 minuto y 2 minutos, siendo el RMSE 6,59 y 7.46 veces mayor que LSTM respectivamente. Con valores de f entre 3 y 10 minutos el factor de incremento que ha sufrido ha sido entre 2.31 y 3.72, hasta llegar a los 20 minutos donde empieza a aproximarse más a LSTM con un factor de 1.37.

Naïve en los dos primeros horizontes temporales se ve superado por la red LSTM, obtenido un factor de 7.08 y 2.03 respectivamente. Después entre los valores de f de 3 y 5 obtiene resultados muy similares. Por último del horizonte temporal de 10 minutos en adelante este método es el que mejor resultados ofrece consiguiendo reduciendo el RMSE obtenido por LSTM, aunque la diferencia sea escasa.

Y por último el método ARIMA, el cual ha obtenido resultados inferiores para $f = 1$, siendo su factor multiplicador de 7.1, pero después se ha aproximado bastante a los resultados obtenidos a LSTM pero sin llegar a superarlos.

Como se puede ver en la Figura 6.3 durante los dos primeros valores de horizonte temporal, la propuesta que utiliza LSTM es la que ha obtenido mejor resultados obtenido notables diferencias. A partir de $f = 3$ las diferencias empiezan a ser menores entre todos los métodos, y según es mayor el horizonte temporal, los resultados obtenidos por los métodos de predicción son mas similares.

Los datos se han obtenido como la media de 3 experimentos por cada día:

- Predicción de radiación en el minuto $12h + f$ a partir de los datos obtenidos entre las 9 a 12

(f en minutos y con los valores indicados en los Cuadros 6.3 y 6.2).

- Predicción de radiación a las $13h + f$ a partir de los datos obtenidos entre las 12 y las 13 horas.
- Predicción de la radiación a las $15h + f$ a partir de los datos obtenidos entre 13 y las 15 horas.

Se ha comprobado que los datos no varían sensiblemente al considerar como conjunto de entrenamiento valores superiores a la hora. Es decir, no hay variación significativa entre considerar predicciones a partir de datos acumulados de la hora anterior y considerando, por ejemplo, las 3 horas anteriores. Esto es positivo porque indica que se pueden comenzar a predecir valores a partir de las 10h de cada día. En cambio, con datos de menos de una hora sí se aprecian aumentos significativos del error.

En resumen, los cuadros muestra la media de un total de 365 (días) x 3 (predicciones por día) x 6 (estaciones meteorológicas) = 6570 tests.

Para comprobar si los datos de los cuadros suponen diferencias significativas se ha llevado a cabo la prueba de los rangos con signo de Wilcoxon[21], utilizada para comparar el rango medio de dos muestras relacionadas y determinar si existen diferencias entre ellas, comprobando que, en efecto, las diferencias de la media de RMSE entre LSTM y cada una de las otras técnicas de predicción muestra diferencias estadísticamente significativas.

Capítulo 7

Conclusiones

Para optimizar el aprovechamiento de la energía solar es importante disponer de una estimación precisa de la energía que se producirá. En este Trabajo de fin de Máster se ha considerado el problema de la predicción de radiación directa a muy corto plazo sobre un conjunto de datos de radiación históricos recopilados en la Plataforma Solar de Almería (PSA-CIEMAT [1]) durante un periodo de tiempo de un año utilizando técnicas de predicción basadas en aprendizaje automático. Para no alterar los procesos siguientes en primer lugar se ha realizado un preprocesado de los datos. También se han utilizado diferentes implementaciones de modelo de cielo claro para reducir el error en la simulación de cielo despejado. Para comprobar si los datos admiten algún tipo de segmentación se han utilizado técnicas de clustering sobre el conjunto de datos, obteniendo como mejor resultado una separación en dos grupos. Se ha conseguido comprobar claramente que las dos distinciones que proporcionaban cada grupo se corresponden a los estados de cielo nuboso y cielo despejado. Se ha comprobado que, mediante la utilización de redes LSTM, es posible mejorar, aunque por un margen estrecho, la predicción basada en la repetición del último valor conocido (predicción Naïve), método simple pero muy efectivo en los primeros minutos, y que sí mejora los resultados de otras técnicas habituales como la predicción basada en la media, incluso de otras más complejas como ARIMA, obteniendo un RMSE hasta 7 veces menor en horizontes temporales reducidos. Gracias a esta predicción a corto plazo se podrán obtener estimaciones de la radiación solar y así realizar una mejor gestión de la energía.
Los resultados obtenidos en este Trabajo de fin de Máster se han presentado al congreso "VI Jornadas en Cloud Computing y Big Data"JCC&BD 2018 (http://jcc.info.unlp.edu.ar/) y a la revista "Journal of Computer Science & Technology (JCS&T)".

[1] http://www.psa.es

Conclusions

To optimize the use of solar energy it is important to dispose an accurate estimate of the energy will produce. In this Master thesis, the problem of predicting very short-term direct radiation over a specific data set collected in the Solar Platform of Almería (PSA-CIEMAT [2]) during a period of one year has been considered using predictive techniques based on machine learning. In order to not alter the following processes, first of all, the data has been preprocessed. Different implementations of the clear sky model have also been used to reduce the error in the clear sky simulation. To check if the data supports some type of segmentation, clustering techniques have been used on the dataset, obtaining the best result a separation in two groups. It has getting to clearly verify the two distinctions which provided each group, cloudy and clear. It has been proven that, by using LSTM networks, it is possible to improve, although by a narrow margin, the prediction based on the repetition of the last known value (Naïve forecasts), a simple but very effective method in the first minutes, and which does improve the results of other common techniques such as prediction based on the average, or even more complex ones such as ARIMA, obtaining an RMSE up to 7 times lower in reduced time horizons. Thanks to this short-term prediction, it will be possible to obtain estimates of solar radiation and so achieve a better management of energy.
The results obtained in this end-of-master's work work have been presented to the Congress "VI Jornadas en Cloud Computing y Big Data"JCC&BD 2018 (http://jcc.info.unlp.edu.ar/) and the journal "Journal of Computer Science & Technology (JCS&T)".

[2] http://www.psa.es

Capítulo 8

Futuras líneas de investigación

Este estudio enmarcado en un Trabajo de Fin de Máster ha dejado diferentes líneas abiertas por las que se puede continuar la investigación, habiendo dejado asentando principalmente la preparación y filtrado de los datos, además de un trabajo de investigación y análisis de posibles soluciones y posibilidades para enfocar el problema. A continuación se mostrarán cuáles son los frentes que ha dejado este trabajo:

***Utilización de datos procedentes de más ubicaciones geográficas*:** Los datos utilizados en este estudio han sido ofrecidos por el CIEMAT, provenientes de la Plataforma Solar de Almería, es decir, únicamente se ha dispuesto de esa ubicación geográfica. Por tanto, este estudio ha sido centrado en las características propias que dispone dicha zona. Para realizar un estudio más amplio sería necesario disponer de un mayor número de ubicaciones geográficas con meteorologías distintas y así poder realizar un estudio más versátil.

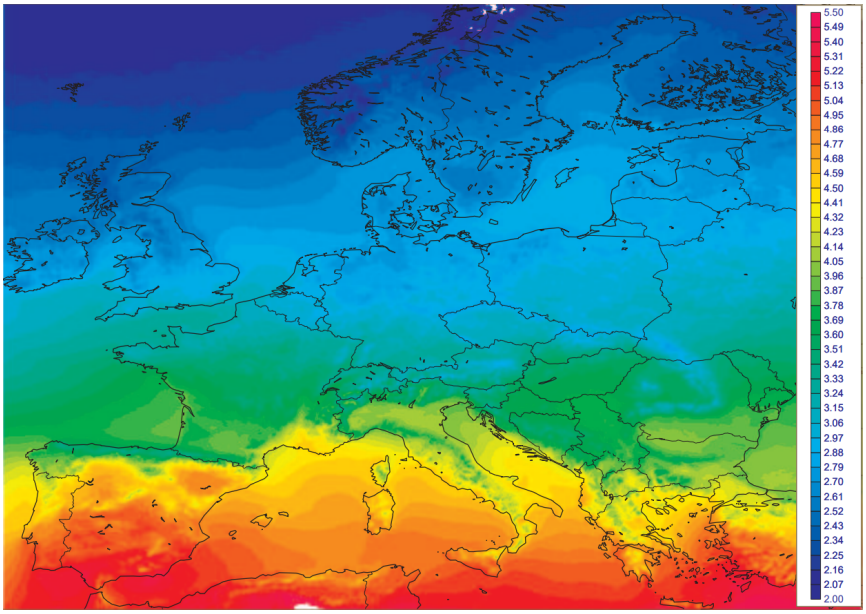


Figura 8.1: Gráfica de la irradiancia Global media en Europa entre 1983 y 2005. Está medido en Kwh m-2 dia-1. Fuente: AEMET

Como se puede ver en la gráfica Figura 8.1, cuanto más al sur se encuentre la ubicación mayor será la radiación, por tanto es importante estudiar exhaustivamente las características concretas de la ubicación.

***Realizar la predicción con datos de varios años de la misma zona geográfica*:** Para realizar la predicción se ha utilizado los datos únicamente del año 2015, por tanto, el estudio realizado puede verse afectado por las particularidades ocurridas en ese periodo de tiempo. Si en su lugar se utilizasen más años, este ruido se vería disminuido pudiendo obtener predicciones más generales para la zona en la que se realiza el estudio.

***Utilizar varios modelos según la época del año*:** Los datos utilizados han sido de todos los meses del año, y con estos se ha generado el modelo de predicción. En la ubicación geográfica en la que se ha realizado el estudio existen grandes variaciones climatologías para las distintas estaciones del año, dificultando así la generación de un único modelo para todo el año.
Por tanto generando un modelo para cada una de las estaciones del año se podrían obtener buenos resultados, consiguiendo un modelo especifico capaz de predecir las características concretas de cada una de las estaciones.
Otra posibilidad es utilizar los meses del año en lugar de utilizar las estaciones como criterio para cada uno de los modelos.

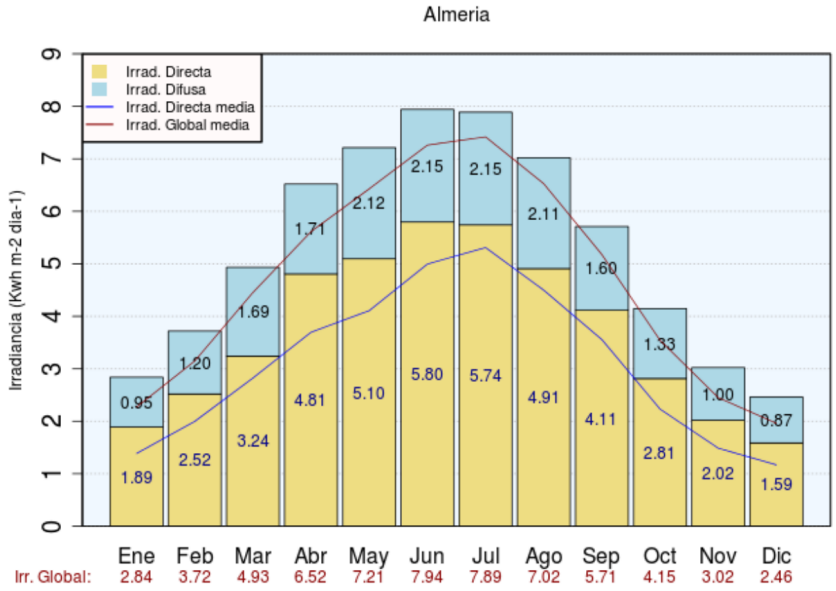


Figura 8.2: Gráfica de la irradiancia de Almería anual distribuida por meses. Fuente: AEMET

***Predicción utilizando de forma conjunta las estaciones solares*:** Este estudio ha tratado cada una de las estaciones como un conjunto aislado, es decir, para predecir el comportamiento de la estación únicamente se ha utilizado sus propios datos. Una forma de obtener mayor información es utilizar para cada estación la información de todas las demás, ya que en numerosas

ocasiones estos datos pueden estar fuertemente relacionados y pueden ayudar a predecir con mayor anticipación el comportamiento que tendrá en los siguientes momentos. Ya que, al estar la estaciones en ubicaciones próximas, si el viento procede de una dirección, las nubes por tanto vendrán de esta misma, y las estaciones situadas anteriormente en esa ubicación detectaran previamente la situación y podremos predecir con mayor anticipación lo que podrá ocurrir en los próximos instantes.

***Normalización mediante otros modelos de cielo claro*:** Como se explicó en el Capítulo 5, en este estudio se realizó la normalización utilizando una librería de python llamada Pysolar y una web. Estos modelos fueron elegidos debido a su fácil acceso y cálculo, pero las aproximaciones que realizan de cielo claro no son del todo precisas, lo que implica un factor de error en el proceso de predicción.

Para poder evitar esto y así obtener unos valores más próximos a los reales, seria conveniente utilizar varios algoritmos de cielo claro más complejos, que tienen en cuenta más parámetros, y seleccionar el que ofrezca un mayor parecido.

Un ejemplo de modelo de cielo claro puede ser el de Bird & Hulstrom[22], ampliamente utilizado y ofrece buenos resultados únicamente utilizando datos meteorológicos: los espesores ópticos de aerosol para longitudes de onde de 500mm y 380mm, el ozono en la columna vertical de la atmósfera y el nivel de vapor de agua.

***Intentar realizar predicciones otras configuraciones de deep learning*:** En este estudio lo que mejor resultado dio fue una red neuronal LSTM. Seria también interesante probar otro tipo de configuraciones de redes neuronales como las CNN (Convolutional Neural Networks). En este tipo de redes neuronales podríamos trabajar con ”imágenes” (matrices de datos) en lugar de con datos aislados. Por tanto, necesitaríamos convertir la información que poseemos en una matriz. Una primera forma de lograrlo es utilizando la posición geográfica poniendo sobre un mapa los puntos en los que están y resto interpolarlos, generando una cuadricula que mantiene la posición real de cada uno. A cada una de estas ”imágenes” se les da una etiqueta de 1 o 0 (nublado o despejado) según si el instante que le precede esta despejado o nublado. Para saber si un punto está nublado o no se puede utilizar la información obtenida al realizar el clustering sobre cuál es el umbral para pertenecer a un estado u otro.

También se podrían utilizar otros tipos de RNN (Recurrent Neural Networks) que son muy utilizadas para modelar datos de una secuencia temporal. Cada nuevo elemento actualiza el estado del modelo aportando información extra. Este tipo de red neuronal puede ser muy conveniente ya que los datos que estamos tratando siguen una serie temporal, por tanto mantener la información temporal puede ayudar a obtener buenos resultado.

***Introducir como feature de entrada otros datos*:** La característica utilizada para realizar la predicción ha sido únicamente la radiación que hubo en esa ubicación en una fecha anterior. Añadiendo más factores al modelo como la altura del sol, excentricidad de la órbita terrestre, concentración de componentes atmosféricos como aerosoles y ozono o el nivel de humedad

en el ambiente podrían ayudar a obtener una estimación más precisa y reducir posibles errores de lectura de algún sensor.
Existen estudios [23] que han utilizado como variables de entrada en la red neuronal la temperatura máxima, la mínima y la radiación extraterrestre. Por tanto, este puede ser un factor determinante a la hora de mejorar los resultados obtenidos en la predicción.
El problema es que no siempre es sencillo obtener todas estas características de la ubicación geográfica a estudiar.

***Realizar la predicción a otros parámetros*:** En este estudio se ha realizado la predicción utilizando la radiación global pero también sería interesante utilizar otros tipos, como por ejemplo la directa y analizar los resultados obtenidos.

Bibliografía

[1] http://www.gigavation.com/?p=156.

[2] T. M. Mitchell *et al.*, "Machine learning. wcb," 1997.

[3] S. Chu, Y. Cui, and N. Liu, "The path towards sustainable energy," *Nature materials*, vol. 16, no. 1, p. 16, 2017.

[4] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging*, vol. 16, no. 4, p. 049901, 2007.

[5] https://lilianweng.github.io/lil-log/2017/06/21/an-overview-of-deep-learning.html.

[6] https://deeplearning4j.org/neuralnet-overview.

[7] A. Ng, "Nuts and bolts of applying deep learning," 2016.

[8] https://www.researchgate.net/.

[9] https://www.wikipedia.org/.

[10] D. Renné, "Status of task 36 solar resource knowledge management under the iea solar heating and cooling programme," tech. rep., NREL (National Renewable Energy Laboratory (NREL), Golden, CO (United States)), 2009.

[11] A. Hammer, D. Heinemann, E. Lorenz, and B. Lückehe, "Short-term forecasting of solar radiation: a statistical approach using satellite data," *Solar Energy*, vol. 67, no. 1-3, pp. 139–150, 1999.

[12] J. L. Bosch and J. Kleissl, "Cloud motion vectors from a network of ground sensors in a solar power plant," *Solar Energy*, vol. 95, pp. 13–20, 2013.

[13] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342–349, 2009.

[14] L. Martín, L. F. Zarzalejo, J. Polo, A. Navarro, R. Marchante, and M. Cony, "Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning," *Solar Energy*, vol. 84, no. 10, pp. 1772–1781, 2010.

[15] C. Paoli, C. Voyant, M. Muselli, and M.-L. Nivet, "Forecasting of preprocessed daily solar radiation time series using neural networks," *Solar Energy*, vol. 84, no. 12, pp. 2146–2160, 2010.

[16] A. Mellit and A. M. Pavan, "A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected pv plant at trieste, italy," *Solar Energy*, vol. 84, no. 5, pp. 807–821, 2010.

[17] M. Bou-Rabee, S. A. Sulaiman, M. S. Saleh, and S. Marafi, "Using artificial neural networks to estimate solar radiation in kuwait," *Renewable and Sustainable Energy Reviews*, vol. 72, pp. 434–438, 2017.

[18] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2014.

[19] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[20] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[21] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[22] R. E. Bird and R. L. Hulstrom, "Simplified clear sky model for direct and diffuse insolation on horizontal surfaces," tech. rep., Solar Energy Research Inst., Golden, CO (USA), 1981.

[23] A. K. Yadav and S. Chandel, "Solar radiation prediction using artificial neural network techniques: A review," *Renewable and sustainable energy reviews*, vol. 33, pp. 772–781, 2014.

[24] S. Cros, O. Liandrat, N. Sébastien, N. Schmutz, and C. Voyant, "Clear sky models assessment for an operational pv production forecasting solution," in *28th European Photovoltaic Solar Energy Conference and Exhibition*, pp. 5BV–4, 2013.

[25] A. J. Serrano, E. Soria, and J. Martín, "Redes neuronales artificiales," *Universidad de Valencia (Escuela Técnica Superior Ingeniería, Departamento de Ingeniería Electrónica): Valencia, España*, 2009.